Hyperspectral Image Classification Based on 3-D Octave Convolution With Spatial–Spectral Attention Network

Xu Tang, *Member, IEEE*, Fanbo Meng, *Graduate Student Member, IEEE*, Xiangrong Zhang, *Senior Member, IEEE*, Yiu-Ming Cheung, *Fellow, IEEE*, Jingjing Ma, *Member, IEEE*, Fang Liu, *Member, IEEE*, and Licheng Jiao, *Fellow, IEEE*

***Abstract*— In recent years, with the development of deep learning (DL), the hyperspectral image (HSI) classification methods based on DL have shown superior performance. Although these DL-based methods have great successes, there is still room to improve their ability to explore spatial–spectral information. In this article, we propose a 3-D octave convolution with the spatial–spectral attention network (3DOC-SSAN) to capture discriminative spatial–spectral features for the classification of HSIs. Especially, we first extend the octave convolution model using 3-D convolution, namely, a 3-D octave convolution model (3D-OCM), in which four 3-D octave convolution blocks are combined to capture spatial–spectral features from HSIs. Not only the spatial information can be mined deeply from the high- and low-frequency aspects but also the spectral information can be taken into account by our 3D-OCM. Second, we introduce two attention models from spatial and spectral dimensions to highlight the important spatial areas and specific spectral bands that consist of significant information for the classification tasks. Finally, in order to integrate spatial and spectral information, we design an information complement model to transmit important information between spatial and spectral attention features. Through the information complement model, the beneficial parts of spatial and spectral attention features for the classification tasks can be fully utilized. Comparing with several existing popular classifiers, our proposed method can achieve competitive performance on four benchmark data sets.**

***Index Terms*— Attention mechanism, deep learning (DL), hyperspectral image (HSI) classification, information complement, spatial–spectral features.**

Manuscript received June 7, 2020; accepted June 22, 2020. Date of publication July 14, 2020; date of current version February 25, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 61801351, Grant 61802190, Grant 61772400, Grant 61672444, and Grant 61272366, in part by the Key Laboratory of National Defense Science and Technology Foundation Project under Grant 6142113180302, in part by the China Post-Doctoral Science Foundation Funded Project under Grant 2017M620441, in part by Xidian University New Teacher Innovation Fund Project under Grant XJS18032, in part by Hong Kong Baptist University (HKBU), Research Committee, Initiation Grant–Faculty Niche Research Areas (IG-FNRA) 2018/19 under Grant RC-FNRA-IG/18-19/SCI/03, in part by the Innovation and Technology Fund of Innovation and Technology Commission of the Government of the Hong Kong SAR under Grant ITS/339/18, in part by the Faculty Research Grant of HKBU under Project FRG2/17-18/082, and in part by the Shenzhen Science, Technology and Innovation Commission (SZSTI) under Grant JCYJ20160531194006833. *(Corresponding authors: Xu Tang; Yiu-Ming Cheung.)*

Xu Tang is with the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, School of Artificial Intelligence, Xidian University, Xi'an 710071, China, and also with the Department of Computer Science, Hong Kong Baptist University, Hong Kong (e-mail: tangxu128@gmail.com).

Fanbo Meng, Xiangrong Zhang, Jingjing Ma, and Licheng Jiao are with the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, School of Artificial Intelligence, Xidian University, Xi'an 710071, China.

Yiu-Ming Cheung is with the Department of Computer Science, Hong Kong Baptist University, Hong Kong (e-mail: ymc@comp.hkbu.edu.hk).

Fang Liu is with the Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China.

Color versions of one or more of the figures in this article are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TGRS.2020.3005431

I. Introduction

As an important product of remote sensing image, hyperspectral images (HSIs) draw the researchers' attention because not only the spatial but also the spectral information of the land-cover targets can be provided at the same time [1]. Due to this specific characteristic, HSIs are widely used in many remote sensing applications, such as forest monitoring and urban management [2]. To accomplish these applications comprehensively, many HSI tasks are developed in recent years, e.g., classification [3], unmixing [4], and anomaly detection [5], [6]. Among these tasks, HSI classification is a fundamental task that focuses on assigning the semantic labels to each pixel within an HSI. Both the classifier design and the pixel-level features extraction/learning are paramount for the classification.

To achieve good classification performance, scholars have made great efforts in the last decades. At the very beginning, many basic machine learning classifiers were chosen to complete the classification, such as decision tree [7], random forest [8], support vector machine (SVM) [9] and its extended variants [10], sparse representation-based classification (SRC) [11], and extreme learning machine (ELM) [12]. Nevertheless, the classification results from the abovementioned pixelwise classifiers cannot reach the satisfactory level as only the spectral features are considered [13]. To address this problem, many spectral–spatial feature-based classification methods have been proposed in the literature. For example,

in order to extract spatial dependencies of HSIs, patchwise feature extraction methods are utilized [14]. Compared with the pixelwise feature extraction approaches, apart from the spectral information, the patchwise algorithms could explore the relationship between pixels as well. In addition, some successful statistic models, such as conditional random field [15] and the Markov random field (MRF) [16], are adopted or improved to capture the spatial and spectral information from HSIs for the classification task. Although these methods improve the classification performance to a certain degree, they heavily depend on the handcrafted features. In other words, the classification maps are satisfactory or are not mainly decided by the low-level features. However, most of the handcrafted features may not be able to represent the complex contents within HSIs, which would limit the classification performance.

Recently, with the development of deep learning (DL), an increasing number of DL-based classification methods have been proposed for HSIs. Due to the strong capacity of feature learning, the existing DL-based methods push the classification performance toward the peak value. For example, the article [17] introduces a deep belief network (DBN) to extract the features and complete the classification at the same time. Similarly, another success network, i.e., stacked autoencoder (SAE) [18], is selected to obtain the classification maps for HSIs. However, the input of the two abovementioned networks is the spectral vector of each pixel, which cannot provide the spatial information. To address this issue, the convolutional neural network (CNN) becomes popular in the HSI community [19]–[21], which utilizes both of the spectral and spatial information to get the classification results. An increasing number of popular CNNs are used to classify the HSIs, such as ResNet [22], CapsNet [23], DenseNet [24], and dual-path networks [25]. ResNet and DenseNet are good at combining the deep and shallow features of HSIs for obtaining the classification results [26], [27]. CapsNet specializes in capturing the relationships between different spectral bands and the resemblance between diverse spatial positions for the HSIs classification tasks [28]. Due to the specific architecture, dual-path networks are apt to explore the spatial and spectral information from HSIs simultaneously for classifying HSIs [29]. Both of them make their contributions to improve the classification performance of HSIs. In addition, many new HSIs classification methods are proposed based on the abovementioned networks [30], [31]. Moreover, more specific networks, such as RNN [32] and LSTM [33], are adopted to regard the continuous spectral bands as the temporal data and analyze them.

In recent years, since the attention mechanism can capture detailed information [34], many methods based on visual attention have been developed to obtain the classification maps for HSIs [35]. Although these DL-based methods have achieved great success, there is still room for improvement. First, some networks are complex, and the number of parameters is huge. Thus, it is hard to train them using the limited labeled HSI data [36]. Second, due to the complexity of HSIs, besides global information, the significant spatial locations and spectral bands are also important for classification. However, this specific information is not fully mined. Finally, most of the acquired spatial and spectral features are independent of each other so that the mutual information between them is missed, which would reduce their contributions to the classification task.

To overcome the abovementioned limitation, we present a new DL-based classification method for HSIs, named 3-D octave convolution with spatial–spectral attention network (3DOC-SSAN)[1]. First, we apply the octave convolution model [37] with a small volume of parameters on HSIs to extract spatial information. Meanwhile, considering the effect of spectral information on the classification task, the extended octave convolution model based on 3-D convolution is developed to capture spectral information. Second, we design two attention mechanisms from both spatial and spectral dimensions [38]. Through adding the attention mechanisms, the interested spatial areas and spectral bands that are beneficial to classification tasks can be highlighted. Finally, in order to integrate the contributions of spatial and spectral features, we generate an information complement method, by which the spatial and spectral features can be fused in a mutual manner. This is beneficial to remain the important parts of different features. The main contributions of this article are summarized as follows.

1) A 3-D octave convolution model (3D-OCM) is developed for learning spatial–spectral features from HSIs simultaneously. In addition, because of the high- and-low-frequency decompositions (from the spatial aspect), the volume of parameters can be reduced compared with the common 3-D convolutional networks.
2) Two attention methods are introduced into our network for capturing the significant spatial areas and exploring the specific spectral bands, which can improve the discrimination of the learned features.
3) An information complement method is proposed to mine the mutual information between the spatial and spectral features. Through this method, the spatial and spectral features can be fused properly for the final classification.
4) Extensive experiments are conducted on four public HSIs. The encouraging results prove that our network is useful for HSI classification tasks.

The remainder of this article is organized as follows. Section II briefly reviews the DL-based HSI classification methods. In Section III, the proposed classification framework is introduced, including the octave convolution model based on 3-D convolution, the attention methods, and the information complement approach. Experimental settings and results counted on four data sets are shown in Section IV. Finally, the conclusion is drawn in Section V.

II. Related Work

With the development of DL, an increasing number of DL-based HSI classification methods are proposed. According to the feature types extracted by the network, we can divide these methods into three groups roughly; they are the algorithms based on spectral, spatial, and spectral–spatial features.

[1]Our source codes are available at https://github.com/smallsmallflypigtang/Hyperspectral-Image-Classification-Based-on-3D-Octave-Convolution-with-Spatial–Spectral-AttentionGithub website.

Due to a large amount of spectral information exists in the continuous spectral bands, the spectral features are important for classification tasks of HSIs. In earlier studies, pixels' spectral vectors are directly input into the networks (e.g., SAE and DBN) to learn the features. Then, some successful classifiers (e.g., SVM) are selected to complete the classification. Based on this architecture, many improved methods are proposed. For example, Liu *et al*. [39] combined DBN and active learning (AL) to design a spectral-based classification method of HSIs. The DBN network is embedded in the AL pipeline. Through the estimation of the representative and uncertainty data, a small number of data can be selected to train the DBN for classifying the HSIs. The article [40] presented a virtual sample-enhanced method to increase the number of samples for solving the problem of insufficient labeled samples and uses 1-D convolution to extract spectral features for classification tasks. Zhan *et al*. [41] proposed a classification method based on 1-D generative adversarial network (GAN). They used unlabeled samples to train the network for obtaining a discriminator and then transformed the trained discriminator into a classification network. Finally, few labeled samples were utilized to fine-tune the network for accomplishing the HSI classification.

Apart from the spectral features, spatial information of HSIs is also important for classification tasks. In order to classify HSIs accurately, the spatial features obtained by the DL network are usually fused with spectral features. It is well known that the classification accuracy can be improved by adding the spatial characteristics to the classification tasks. In the article [42], the principal component analysis (PCA) was performed to reduce the dimensionality of HSIs, and then, the 2-D CNN was used to convolve the dimension-reduced data for extracting the spatial information of HSIs. The combination of PCA and 2-D CNN can obtain spatial features with low computation cost. Jiao *et al*. [43] adopted pretrained full CNN (FCN) to explore multiscale spatial structural information. Then, they combined original spectral features and multiscale spatial features for the classification tasks. Although the multiscale spatial features have much superiority, the issue of spatial resolution reduction would influence the classification results. To overcome this drawback, Niu *et al*. [44] proposed a novel HSI classification framework based on the semantic segmentation idea. They applied the minimum noise fraction (MNF) to reduce dimensions and acquire the pseudolabels of samples. Then, the spatial features at multiple scales were extracted by the DeepLab [45] to complete the classification with SVM. The DeepLab ensures the effectiveness of multiscale spatial features and avoids the reduction of the spatial resolution.

In the abovementioned methods, spatial features or spectral features are always extracted separately for the classification tasks. Recently, using the joint spectral–spatial features to classify HSIs has also got excellent results. Instead of acquiring spectral and spatial features separately, the joint spectral–spatial features can be obtained by DL networks directly. From the structure aspect, HSI is a 3-D cube data. Thus, it is proper to use 3-D convolution to extract the spectral–spatial features jointly. For instance, Yang *et al*. [46] designed a recurrent 3-D CNN method to learn the joint spectral–spatial features through shrinking the patch gradually. The learned features contain both spatial and spectral context relations and alleviate the influence of patch size on classification accuracy. Mou *et al*. [47] developed a deep residual convolution–deconvolution network for HSI classification. They trained a deep residual convolution–deconvolution network with unlabeled samples. Then, the encoder of the network was fine-tuned with a small number of labeled data to complete the spectral–spatial feature learning and classification at the same time. Hang *et al*. [48] created a spectral–spatial cascaded RNN model for the classification tasks of HSIs. They set a gated recurrent unit consists of two layers of RNN to process the joint spectral–spatial features. The redundant information of spectral bands can be reduced by the first-layer RNN, and the second-layer RNN is used to learn complementary features from the different reduced information.

Besides the abovementioned methods, the attention mechanism has received increasing attention in the HSI processing community recently. By adding the visual attention, the important spatial areas and the paramount spectral bands can be highlighted, which are beneficial to improve the classification performance. Sun *et al*. [38] proposed a spectral–spatial attention network to increase the classification accuracy of HSIs. They set spectral and spatial models with 3-D convolution to extract the joint spectral–spatial features. Then, the attention model was embedded between two models to suppress the effects of interfering pixels and capture attention areas in an HSI cube. Mou and Zhu [49] proposed a learnable spectral attention model for classification of HSIs. They produced a spectral gate with a global convolution to exploit the global spectral–spatial context relationship. Then, multiplying the spectral gate with original HSI data to recalibrate spectral information, this can effectively improve the classification results. Haut *et al*. [35] designed a visual attention network of two paths, i.e., trunk path and mask path, for classifying HSIs. The trunk path aims to extract spectral–spatial features, and the mask path focuses on calculating and multiplying the attention mask to the trunk path. Due to the visual attention techniques, the abovementioned methods can improve the discrimination of features and enhance the stability of the model.

III. Proposed Method

The proposed framework of 3DOC-SSAN is illustrated in Fig. 1, which consists of the 3D-OCM, the spatial–spectral attention model (SSAM), and the spatial–spectral information complement model (SSICM). First, the 3D-OCM block is developed to capture the spatial–spectral features $\mathbf{F}^o$ from the HSIs. By combining the octave convolutional network and the 3-D convolution subtly, not only the spatial information but also the spectral information can be explored from the HSIs simultaneously. Second, to improve the discrimination of $\mathbf{F}^o$, the SSAM block is introduced, in which the significant areas within $\mathbf{F}^o$ can be fully explored using the channelwise and spatialwise attention methods. After this operation, two feature maps $\mathbf{A}^{\text{spa}}$ and $\mathbf{A}^{\text{spe}}$ can be obtained, and they would bring much more discriminative information. The information within

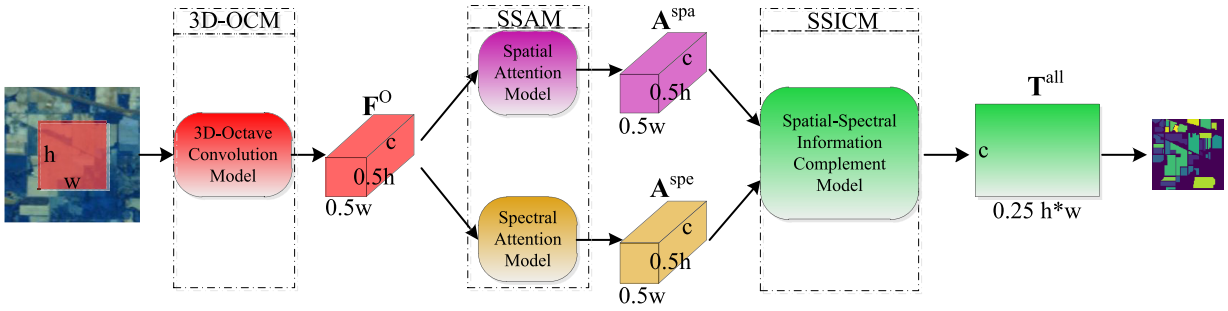


Fig. 1. Flowchart of 3DOC-SSAN.

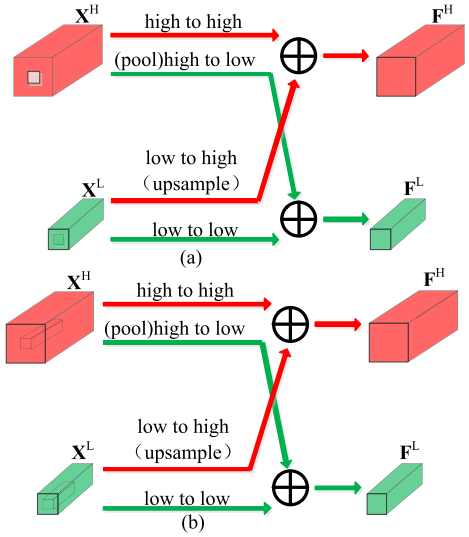


Fig. 2. (a) and (b) Octave convolution block and the 3-D octave convolution block.

$\mathbf{A}^{spa}$ and $\mathbf{A}^{spe}$ is different and complementary. Considering this point, we develop the SSICM block to integrate the contributions of $\mathbf{A}^{spa}$ and $\mathbf{A}^{spe}$ for the classification results in a mutual learning manner. Through the SSCIM block, the important information within $\mathbf{A}^{spa}$ and $\mathbf{A}^{spe}$ can be remained, and the redundant information within $\mathbf{A}^{spa}$ and $\mathbf{A}^{spe}$ can be removed. Note that similar to the existing methods [50]–[52], we pick up an image patch centered at each pixel rather than the individual pixel to build our classification model. Now, we introduce each model of the 3DOC-SSAN.

A. 3D-OCM

Before explaining 3D-OCM, we introduce the octave convolution (Oct-Conv) first. The Oct-Conv block was proposed in the literature [37], and the basic flowchart is shown in Fig. 2(a).

The Oct-Conv block is developed for the natural image originally, and it assumes that a natural image can be decomposed into low and high frequencies, which could represent global structures and local fine details, respectively. Thus, a two-branch convolutional framework is developed in the Oct-Conv block to capture global and local information. Due to the low-frequency branch, the number of parameters can be decreased, the channelwise redundancy can be reduced, and the receptive field is enlarged [37], [53]. In addition, to ensure the integrity of the information, a communication mechanism is established between two frequencies to compliment the diverse information corresponding to high and low parts mutually.

In detail, suppose that the input and output data of an Oct-Conv block are $\mathbf{X} = \{\mathbf{X}^H, \mathbf{X}^L\}$ and $\mathbf{F} = \{\mathbf{F}^H, \mathbf{F}^L\}$, where the superscript H and L indicate the high and low frequencies, respectively. The Oct-Conv model defines $\mathbf{F}^H = \mathbf{F}^{H\to H} + \mathbf{F}^{L\to H}$ and $\mathbf{F}^L = \mathbf{F}^{H\to L} + \mathbf{F}^{L\to L}$, where $\mathbf{F}^{H\to H}$ and $\mathbf{F}^{L\to L}$ mean the intrafrequency transition, while $\mathbf{F}^{H\to L}$ and $\mathbf{F}^{L\to H}$ denote the interfrequency update. In order to accomplish the information update and interaction mentioned earlier, the weights of the Oct-Conv block $\mathbf{W}$ should be divided into two parts $[\mathbf{W}^H, \mathbf{W}^L]$ as well. Furthermore, each element can be partitioned into the intra- and inter-frequency components, e.g., $\mathbf{W}^H = [\mathbf{W}^{H\to H}, \mathbf{W}^{L\to H}]$ and $\mathbf{W}^L = [\mathbf{W}^{H\to L}, \mathbf{W}^{L\to L}]$. Thus, $\mathbf{F}^H$ and $\mathbf{F}^L$ can be calculated by the following equations:

$$\begin{aligned} F^H &= F^{H\to H} + F^{L\to H} \\ &= \sum \left(W^H\right)^{\mathrm{T}} X \\ &= \sum \left(W^{H\to H}\right)^{\mathrm{T}} X^H + \mathrm{upsample}\left(\sum \left(W^{L\to H}\right)^{\mathrm{T}} X^L\right) \quad (1) \\ F^L &= F^{H\to L} + F^{L\to L} \\ &= \sum \left(W^L\right)^{\mathrm{T}} X \\ &= \sum \left(W^{H\to L}\right)^{\mathrm{T}} \mathrm{pool}\left(X^H\right) + \sum \left(W^{L\to L}\right)^{\mathrm{T}} X^L \quad (2) \end{aligned}$$

where T represents the transposition of the weights, and upsample and pool stand for the upsampling and average pooling operation, respectively. Although the Oct-Conv block has obvious superiorities, it is improper to apply it to the HSIs classification tasks directly. Due to the characteristic of HSIs, apart from the spatial regions, the continuous spectral bands should be taken into account for the classification task. In the Oct-Conv block, all the convolution operations use 2-D convolution. Although the 2-D convolution can extract spatial and spectral features from HSIs at the same time, the consistency of different spectral bands cannot be explored by the 2-D convolution as it only works on the spatial dimension. Comparatively speaking, due to the cube structure, the 3-D convolution can not only work on the spatial dimension but also convolve several continuous spectral bands at one time. Thus, the 3-D convolution can explore more comprehensive spectral–spatial information from HSIs [46]. Considering the

above-discussed properties, we expand the traditional Oct-Conv to the 3-D version, named 3DOct-Conv. Its structure is shown in Fig. 2(b). The 3DOct-Conv block takes 3-D convolution to convolve spatial regions and spectral bands. Thus, the spatial and spectral contexts of HSIs can be extracted simultaneously.

To illustrate the difference between 3DOct-Conv and Oct-Conv clearly, we introduce 3-D convolution and 2-D convolution here. The 2-D convolution operator is formulated as

$$v_{\text{out}}^{xy} = \sum_{p=0}^{D_i-1} \sum_{q=0}^{E_i-1} w^{pq} v_{\text{in}}^{(x+p)(y+q)} \tag{3}$$

where v_{in}^{xy} and v_{out}^{xy} stand for the input and output at position (x, y) of the feature maps, w^{pq} is the value at the position (p, q) of the convolution kernel, and D_i and E_i are the width and height of the kernel. Compared with 2-D convolution, the kernel of the 3-D convolution adds one dimension, and the 3-D convolution is formulated as

$$v_{\text{out}}^{xyz} = \sum_{p=0}^{D_i-1} \sum_{q=0}^{E_i-1} \sum_{r=0}^{K_i-1} w^{pqr} v_{\text{in}}^{(x+p)(y+q)(z+r)} \tag{4}$$

where v_{in}^{xyz} and v_{out}^{xyz} represent the input and output at position (x, y, z) of the feature maps, w^{pqr} is the value at the position (p, q, r) of the convolutional kernel, and K_i is the size of the 3-D kernel along to the z dimension. Compared with the 2-D convolution, 3-D convolutional kernel could slip x, y, and z dimensions at the same time. Thus, applying 3-D convolution to HSIs, both spatial and spectral information can be learned at the same time.

Based on the 3DOct-Conv block, we construct 3D-OCM, and its structure is shown in Fig. 3. 3D-OCM involves four 3DOct-Conv blocks, an average pooling operation layer, and an upsampling operation layer. The original HSI is regarded as a high frequency because it contains the complete spatial and spectral information. Therefore, in the first 3DOct-Conv block, only high-frequency data $\mathbf{X}^H$ are input to our 3D-OCM network. After the second 3DOct-Conv block, the pooling operation is used to downsample the high-frequency feature maps $\mathbf{F}_2^H$. Then, the downsampled results and the low-frequency feature maps $\mathbf{F}_2^L$ are combined to be the input $\mathbf{F}^{\text{pool}}$ of the third 3DOct-Conv block. This can preserve significant features and reduce the feature maps' dimension of HSIs. Like the first 3DOct-Conv block, $\mathbf{F}^{\text{pool}}$ represents the high frequencies. Since the input of the network is considered to be the high frequency with local fine details, the output $\mathbf{F}^o$ is also regarded as the high frequency. Moreover, in order to ensure the integrity of information, we need to fuse low-frequency features maps $\mathbf{F}_4^L$ into $\mathbf{F}_4^H$. Thus, an upsampling operation is used for $\mathbf{F}_4^L$ that from the last 3DOct-Conv block.

B. SSAM

Although 3D-OCM can capture spatial–spectral features, the discrimination of features needs to be improved. To this end, we introduce the SSAM network to capture the discriminable information from spatial and spectral aspects, respectively. The SSAM network consists of two parts: spatial attention model and spectral attention model. The spatial dependencies between two positions of feature maps and the relation of spatial context of HSIs can be captured by the spatial attention model, while the spectral dependencies between two bands of feature maps and the emphasized informative bands of HSIs can be obtained by the spectral attention model.

1) Spatial Attention Model: The basic flowchart is shown in Fig. 4(a). Let $\mathbf{F}^o \in \mathbb{R}^{h\times w\times c}$ denote the input of the spatial attention model, where h, w, and c indicate the height, width, and band of $\mathbf{F}^o$. First, $\mathbf{F}^{\text{spa}C} \in \mathbb{R}^{h\times w\times c}$ can be obtained by taking $\mathbf{F}^o$ through a convolution layer. Then, $\mathbf{F}^{\text{spa}C}$ is reshaped to $\mathbf{F}^{\text{spa}S} \in \mathbb{R}^{n\times c}$, where $n = w \times h$. Second, in order to acquire spatial attention map $\mathbf{M}^{\text{spa}} \in \mathbb{R}^{n\times n}$, a softmax layer is applied to the product of the matrix $\mathbf{F}^{\text{spa}S}$ and $\mathbf{F}^{\text{spa}T}$, where $\mathbf{F}^{\text{spa}T}$ is transposed by $\mathbf{F}^{\text{spa}S}$. The formulation of $\mathbf{M}^{\text{spa}}$ is given as

$$\mathbf{M}_{ji}^{\text{spa}} = \frac{\exp\left(\mathbf{F}_i^{\text{spa}S} \otimes \mathbf{F}_j^{\text{spa}T}\right)}{\sum_{i=1}^{n} \exp\left(\mathbf{F}_i^{\text{spa}S} \otimes \mathbf{F}_j^{\text{spa}T}\right)} \tag{5}$$

where $\mathbf{M}_{ji}^{\text{spa}}$ represents the ith position's relationship with the jth position, and $\otimes$ denotes the matrix multiplication operation. After that, we multiply $\mathbf{M}^{\text{spa}}$ with $\mathbf{F}^{\text{spa}S}$ and reshape the result to $\mathbb{R}^{h\times w\times c}$. Then, we add the result with $\mathbf{F}^o$ to obtain the output $\mathbf{A}^{\text{spa}A} \in \mathbb{R}^{h\times w\times c}$. The formulation of $\mathbf{A}^{\text{spa}A}$ is

$$\mathbf{A}^{\text{spa}A} = \text{reshape}(\mathbf{M}^{\text{spa}} \otimes \mathbf{F}^{\text{spa}S}) + \mathbf{F}^o \tag{6}$$

where reshape$(\cdot)$ represents the reshaping operation.

The output $\mathbf{A}^{\text{spa}A}$ contains the spatial features of all the positions and highlights the information of important spatial locations. In order to enhance the nonlinearity of $\mathbf{A}^{\text{spa}}$, we take $\mathbf{A}^{\text{spa}A}$ through a convolution layer with the kernel size of 1×1 at the end of the model.

2) Spectral Attention Model: The framework is shown in Fig. 4(b). Similar to the spatial attention model, $\mathbf{F}^o$ is also the input data. First, $\mathbf{F}^o$ is reshaped into $\mathbf{F}^{\text{spe}S} \in \mathbb{R}^{n\times c}$, and then, $\mathbf{F}^{\text{spe}S}$ is transposed into $\mathbf{F}^{\text{spe}T}$. Second, we apply a softmax layer to the product of the matrix $\mathbf{F}^{\text{spe}T}$ and $\mathbf{F}^{\text{spe}S}$ for generating spectral attention map $\mathbf{M}^{\text{spe}} \in \mathbb{R}^{c\times c}$. The formulation of $\mathbf{M}^{\text{spe}}$ is

$$\mathbf{M}_{ji}^{\text{spe}} = \frac{\exp\left(\mathbf{F}_i^{\text{spe}T} \otimes \mathbf{F}_j^{\text{spe}S}\right)}{\sum_{i=1}^{c} \exp\left(\mathbf{F}_i^{\text{spe}T} \otimes \mathbf{F}_j^{\text{spe}S}\right)} \tag{7}$$

where $\mathbf{M}_{ji}^{\text{spe}}$ indicates the spectral relationship of ith and jth bands. In order to obtain the improved feature map $\mathbf{A}^{\text{spe}A} \in \mathbb{R}^{h\times w\times c}$, the attention map and the feature maps should be combined together. Thus, we reshape the product of $\mathbf{M}^{\text{spe}}$ and $\mathbf{F}^{\text{spe}S}$ to $\mathbb{R}^{h\times w\times c}$ and add $\mathbf{F}^o$ to the result. The contents discussed earlier can be formulated as

$$\mathbf{A}^{\text{spe}A} = \text{reshape}(\mathbf{M}^{\text{spe}} \otimes \mathbf{F}^{\text{spe}S}) + \mathbf{F}^o. \tag{8}$$

The output $\mathbf{A}^{\text{spe}A}$ includes spectral relationships of all the bands and emphasizes informative bands. Finally, $\mathbf{A}^{\text{spe}A}$ is convoluted with a 1×1 convolution kernel to get the output $\mathbf{A}^{\text{spe}}$ that contains the information of all bands.

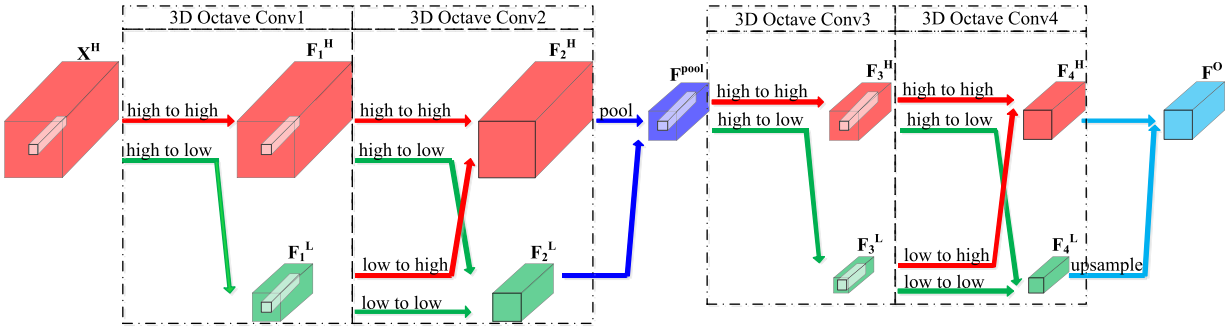


Fig. 3. 3D-OCM.

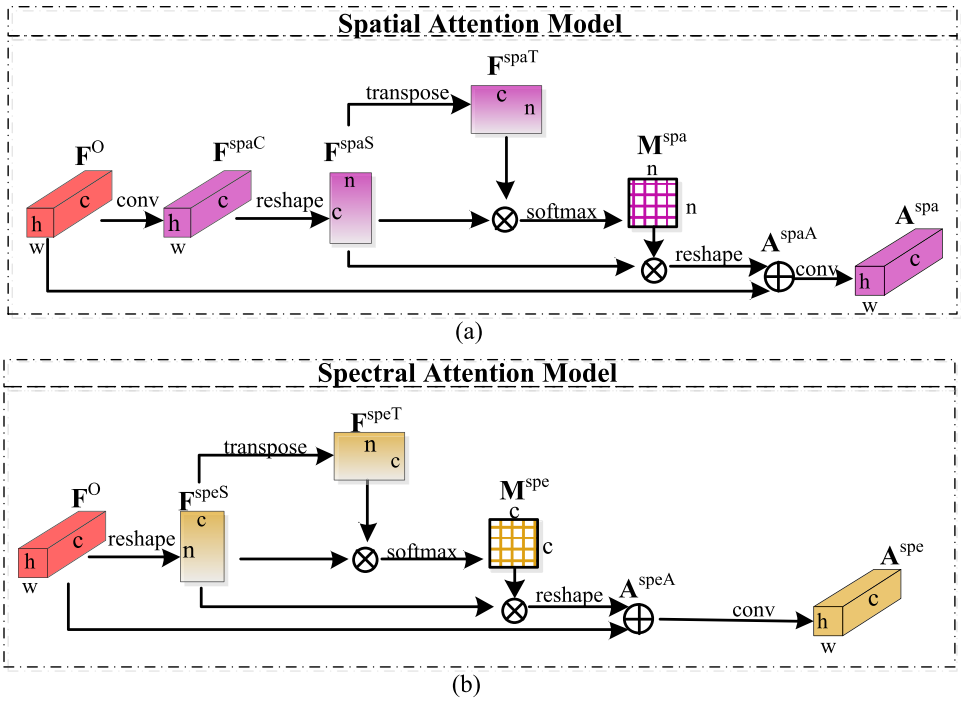


Fig. 4. Two parts of SSAM. (a) and (b) Spatial and spectral attention models.

C. SSICM

Compared with applying spatial or spectral features for the classification task, high accuracy can be obtained by using the fused spatial–spectral features [44], [54], [55]. Here, we design SSICM to establish information flows for transmitting important information between spatial and spectral features so that their contributions can be fully fused. The key to the information flows is to learn a complement matrix. Thus, the complementary spatial features $\mathbf{T}^{\text{spa}}$ contain important spectral information, and the complementary spectral features $\mathbf{T}^{\text{spe}}$ contain detailed spatial information.

For clarity, we illustrate the information flows between the two features. Its structure is shown in Fig. 5. First, the features $\mathbf{A}^{\text{spe}}$ and $\mathbf{A}^{\text{spa}}$ are reshaped into $\mathbf{A}^{\text{speW}} \in \mathbb{R}^{n\times c}$ and $\mathbf{A}^{\text{spaW}} \in \mathbb{R}^{n\times c}$. Then, $\mathbf{A}^{\text{speW}}$ and $\mathbf{A}^{\text{spaW}}$ are transposed to $\mathbf{A}^{\text{speT}}$ and $\mathbf{A}^{\text{spaT}}$. Second, the information transmissions between two kinds of attention features are calculated as follows:

$$C_{\text{spa}\to\text{spe}} = \left[\text{softmax}(\mathbf{A}^{\text{speW}} \otimes \mathbf{A}^{\text{spaT}})\right] \otimes \mathbf{A}^{\text{spaW}} \quad (9)$$

$$C_{\text{spe}\to\text{spa}} = \left[\text{softmax}(\mathbf{A}^{\text{spaW}} \otimes \mathbf{A}^{\text{speT}})\right] \otimes \mathbf{A}^{\text{speW}} \quad (10)$$

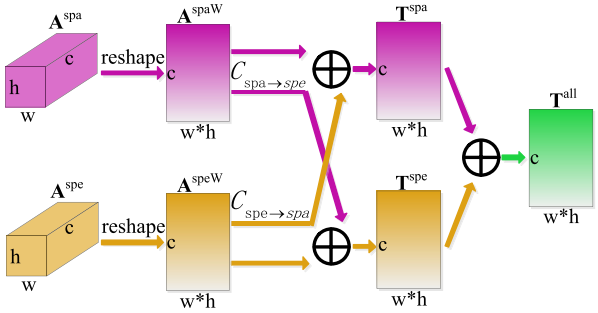
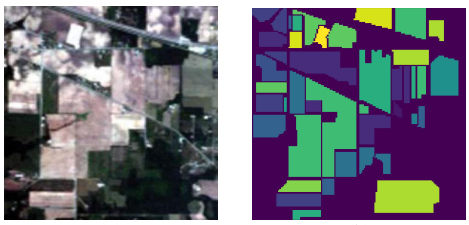


Fig. 5. SSICM.

where $C_{\text{spa}\to\text{spe}}$ donates the information flows from spatial feature to spectral feature and $C_{\text{spe}\to spa}$ is opposite, $\text{softmax}(\mathbf{A}^{\text{spaW}} \otimes \mathbf{A}^{\text{speT}})$ represents spatial complement weight from spatial-to-spectral feature that can stress useful position of the spatial feature, and $softmax(\mathbf{A}^{\text{spaW}} \otimes \mathbf{A}^{\text{speT}})$ represents spectral complement weight from spectral-to-spatial feature that can highlight the detailed spectral bands. By multiplying the spatial complement matrix with $\mathbf{A}^{\text{spaW}}$, the information



(a) (b)

Fig. 6. (a) and (b) Image and labels of the Indian Pines data set.

transmitted from spatial-to-spectral feature can be obtained. In the same way, multiplying the spectral complement matrix with $\mathbf{A}^{\text{spe}W}$ can acquire information which transmitted from spectral-to-spatial feature. In order to integrate spatial information into spectral features, we add $I_{\text{spa}\to spe}$ with $\mathbf{A}^{\text{spe}W}$ to obtain $\mathbf{T}^{\text{spe}}$. Similarly, $\mathbf{T}^{\text{spa}}$ is calculated in the same way. The two formulations are

$$\mathbf{T}^{\text{spe}} = C_{\text{spa}\to\text{spe}} + \mathbf{A}^{\text{spe}W} \tag{11}$$

$$\mathbf{T}^{\text{spa}} = C_{\text{spe}\to\text{spa}} + \mathbf{A}^{\text{spa}W}. \tag{12}$$

D. Optimization Strategy

In order to enhance feature representations of $\mathbf{T}^{\text{spa}}$ and $\mathbf{T}^{\text{spe}}$, we separately use cross-entropy functions for $\mathbf{T}^{\text{spa}}$ and $\mathbf{T}^{\text{spe}}$ to optimize the two branches of the network. At the same time, in order to make all the spatial and spectral information contributed to the classification task, we add $\mathbf{T}^{\text{spa}}$ with $\mathbf{T}^{\text{spe}}$ to get the fusion feature $\mathbf{T}^{\text{all}}$. Similarly, a cross-entropy function is applied to the feature $\mathbf{T}^{\text{all}}$ to optimize the whole network. As three cross-entropy functions are used to optimize the network, the loss function of 3DOC-SSAN is the sum of these three cross-entropy functions. Finally, due to the fusion feature $\mathbf{T}^{\text{all}}$ consisting of all important location information and emphasized spectral bands, the classification result obtained by $\mathbf{T}^{\text{all}}$ is regarded as the output of 3DOC-SSAN.

IV. Experiments

A. Data Description

Four popular HSI data sets, i.e., Indian Pines, Pavia University scene, Botswana, and Houston, are utilized to evaluate the performance of the proposed method.

The Indian Pines data set was collected by the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) sensor over the Indian Pines test site in Northwestern Indiana. It consists of 145 × 145 pixels and 224 spectral reflectance bands with the wavelength range of 0.4–2.5 μm. After removing water absorption bands, there are 200 spectral bands remained. A total of 10 249 pixels are manually labeled and divided into 16 land covers. The false-color image and its corresponding ground-truth map of the Indian Pines data set are shown in Fig. 6, while the number of labeled pixels of each land cover is shown in Table I.

The University of Pavia data set was acquired by the Reflective Optics Imaging Spectrometer (ROSIS) sensor during a flight campaign over Pavia, Northern Italy. The number of original spectral bands with a range of 430–860 nm is 115.

TABLE I

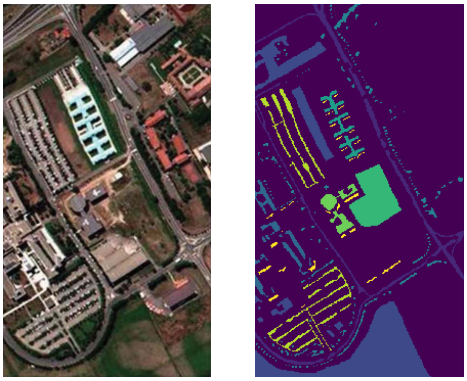
Number of Classes and Pixels of the Indian Pines Data Set

Class		Number		
Code	Name	Train	Test	Total
1	Alfalfa	30	16	46
2	Corn-notill	150	1278	1428
3	Corn-mintill	150	680	830
4	Corn	100	137	237
5	Grass-pasture	150	333	483
6	Grass-trees	150	580	730
7	Grass-pasture-mowed	20	8	28
8	Hay-windrowed	150	328	478
9	Oats	15	5	20
10	Soybean-notill	150	822	972
11	Soybean-mintill	150	2305	2455
12	Soybean-clean	150	443	593
13	Wheat	150	55	205
14	Woods	150	1115	1265
15	Buildings-Grass-Trees-Drives	50	336	386
16	Stone-Steel-Towers	50	43	93
total		1765	8484	10249

TABLE II

Number of Classes and Pixels of the University of Pavia Data Set

Class		Number		
Code	Name	Train	Test	Total
1	Asphalt	548	6083	6631
2	Meadows	540	18109	18649
3	Gravel	392	1707	2099
4	Trees	542	2522	3064
5	Painted Metal Sheets	256	1089	1345
6	Bare Soil	532	4497	5029
7	Bitumen	375	955	1330
8	Self-Blocking Bricks	514	3168	3682
9	Shadows	231	716	947
total		3930	38846	42776



(a) (b)

Fig. 7. (a) and (b) Image and labels of the University of Pavia data set.

After eliminating 12 noisy bands, 103 bands have remained for the classification task. The size of the University of Pavia data set is 610 × 340, and the geometric resolution of pixels is 1.3 m. There are nine semantic categories defined in the University of Pavia data set. The false-color image of the University of Pavia data set and its ground-truth map are shown in Fig. 7, while the numbers of labeled pixels corresponding to different categories are summarized in Table II.

The Botswana data set was collected by the NASA EO-1 Hyperion sensor over the Okavango Delta, Botswana,

TABLE III
NUMBER OF CLASSES AND PIXELS OF THE BOTSWANA DATA SET

Class		Number		
Code	Name	Train	Test	Total
1	Water	30	240	270
2	Hippo grass	30	71	101
3	Floodplain grasses1	30	221	251
4	Floodplain grasses2	30	185	215
5	Reeds1	30	239	269
6	Riparian	30	239	269
7	Firescar2	30	229	259
8	Island interior	30	173	203
9	Acacia woodlands	30	284	314
10	Acacia shrublands	30	218	248
11	Acacia grasslands	30	275	305
12	Short mopane	30	151	181
13	Mixed mopane	30	238	268
14	Exposed soils	30	65	95
total		420	2828	3248

TABLE IV
NUMBER OF CLASSES AND PIXELS OF THE HOUSTON DATA SET

Class		Number		
Code	Name	Train	Test	Total
1	Grass Health	198	1053	1251
2	Grass Stressed	190	1064	1254
3	Grass Synthetic	192	505	697
4	Tree	188	1056	1244
5	Soil	186	1056	1242
6	Water	182	143	325
7	Residential	196	1072	1268
8	Commercial	191	1053	1244
9	Road	193	1059	1252
10	Highway	191	1036	1227
11	Railway	181	1054	1235
12	Parking Lot 1	192	1041	1233
13	Parking Lot 2	184	285	469
14	Tennis Cou	181	247	428
15	Running Track	187	473	660
total		2832	12197	15029

in 2001–2004. It has 242 bands covering the 400–2500-nm portion of the spectrum. After removing uncalibrated and noisy bands, there are 145 spectral bands used for the land-cover classification. The size of the Botswana data is 1476×256, and it has 14 identified classes. The false-color image of the Botswana data set, and its ground-truth maps are shown in Fig. 8, while the numbers of labeled pixels corresponding to different categories are summarized in Table III.

The Houston data set was acquired by the ITRES-CASI 1500 sensor over the University of Houston campus and its neighboring urban area on June 23, 2012. There are 144 spectral bands in this HSI. Its size is 349×1905, and its spatial resolution is 2.5 m. It was published in the 2013 IEEE Geoscience and Remote Sensing Society (GRSS) data fusion contest. The labeled data are grouped into 15 land-cover classes. Also, the training data and testing data of this HSI are defined apart, which enhances the difficulty of the classification. The details of the semantic classes and the numbers of the training and testing data corresponding to each class can be found in Table IV. The false-color image of the Houston data set and its ground-truth map are shown in Fig. 9.

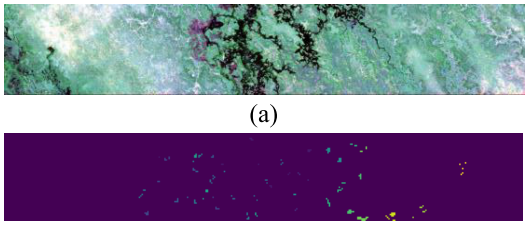


Fig. 8. (a) and (b) Image and labels of the Botswana data set.

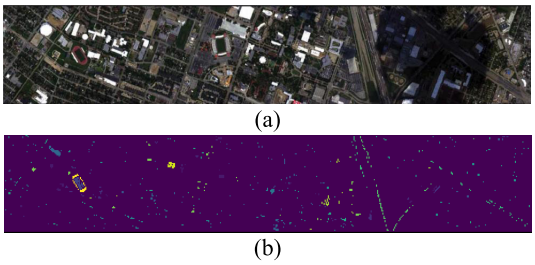


Fig. 9. (a) and (b) Image and labels of the Houston data set.

TABLE V
NETWORK PARAMETERS OF 3DOC-SSAN IN THE INDIAN PINES DATA SET

Model	Operation	Filter	Configuration
3D-OCM	3DOct-Conv1	(5, 3, 3), 24	stride 1,padding 1,bn, relu
	3DOct-Conv2	(5, 3, 3), 48	stride 1,padding 1,bn, relu
	3DOct-Conv3	(5, 3, 3), 24	stride 1,padding 1,bn, relu
	3DOct-Conv4	(5, 3, 3), 1	stride 1,padding 1,bn, relu
SSAM	spa-conv1	(3, 3), 200	stride 1,padding 1,bn, relu
	spa-conv2	(1, 1), 200	stride 1,padding 1,bn, relu
	spe-conv	(1, 1), 200	stride 1,padding 1,bn, relu

B. Experimental Settings and Assessment Criteria

To accomplish the classification task, we randomly select a few numbers of labeled pixels to construct the training data, and the rest of the labeled data are used as the testing data. For different HSIs, the numbers of training and testing pixels are displayed in Tables I–IV. Due to the structure of our 3DOC-SSAN, we use the image patches rather than the pixels to be the input, and the patch size is 13×13 in the following experiments unless otherwise stated. The influence of different patch sizes will be discussed in Section IV-F. In this article, we select the Adam algorithm to train the proposed network. In addition, the learning rate is fixed as 1×10^{-4}, the batch size is equal to 16, and the epochs are set to be 300, 100, 200, and 300 for the Indian Pines, the University of Pavia, Botswana, and the Houston data sets. Besides the above-discussed issues, some key parameters of our model are summarized in Table V.

To evaluate the performance of our method, three assessment criteria are chosen: overall accuracy (OA), average accuracy (AA), and kappa coefficient (K). OA is the ratio of the number of correctly predicted test samples to the number of all test samples. AA is the mean of classification accuracies in all categories. K is defined to measure the consistency between the classification results and ground truth. The higher

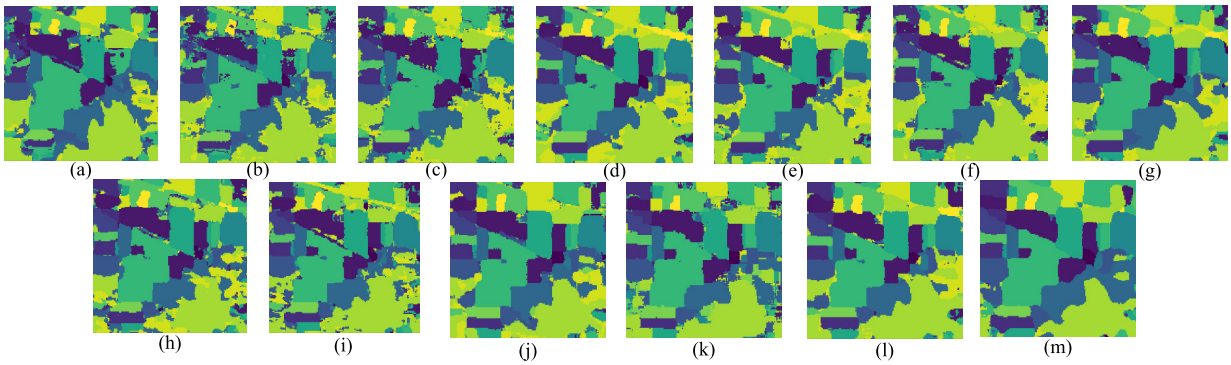


Fig. 10. Classification maps of different methods on the Indian Pines data set. (a) SVM (85.25%). (b) RF-200 (87.48%). (c) Conv-Deconv-Net (95.34%). (d) 2D-CNN (97.02%). (e) C-2D-CNN (96.24%). (f) 3D-CNN (93.42%). (g) SpecAttenNet (98.50%). (h) CAPSNET (99.01%). (i) DPRESNET (98.99%). (j) SSRN (98.98%). (k) DRNN (96.86%). (l) 2DOC-SSAN (98.89%). (m) 3DOC-SSAN (99.14%).

the values of OA, AA, and K, the better the classification results.

C. Compare With Other Methods

In order to verify the effectiveness of 3DOC-SSAN, we select different methods for the comparison, including the traditional machine learning algorithms and the DL-based approaches. The traditional machine learning algorithms are SVM [9] and random forest [8] with 200 decision trees (RF-200). The DL-based approaches are Conv-Deconv-Net [47], 2D-CNN [46], recurrent 2-D CNN (C-2D-CNN) [46], 3D-CNN [40], spectral attention network (SpecAttenNet) [49], CAPSNET [28], DPRESNET [26], SSRN [52], and DRNN [51]. In addition, we change 3-D convolution into a common 2-D convolution to study our 3DOC-SSAN model deeply, and this model is recorded 2DOC-SSAN. For the sake of fairness, all of the comparisons are conducted under the same conditions as 3DOC-SSAN, including parameter settings and data preprocessing.

1) Analysis of Indian Pines Data Set: The visual and numerical classification results of different methods counted on the Indian Pines data set are shown in Fig. 10 and Table VI. From the observation of Fig. 10, we can easily find that the classification map obtained by our 3DOC-SSAN model is clearer than that of the compared methods. Not only the regional consistency but also the boundaries between regions are classified well. For the numerical results (displayed in Table VI), the proposed model achieves the best performance from the overall aspect. Compared with the other methods, the OA scores' enhancements obtained by our model are 13.89% (SVM), 11.66% (RF-200), 3.80% (Conv-Deconv-Net), 2.12% (2D-CNN), 2.90% (C-2D-CNN), 5.72% (3D-CNN), 0.64% (SpecAttenNet), 0.13% (SAPSNET), 0.15% (DPRESNET), 0.16% (SSRN), 2.28% (DRNN), and 0.25% (2DOC-SSAN). The AA scores' improvements are 6.98% (SVM), 6.81% (RF-200), 1.81% (Conv-Deconv-Net), 1.62% (2D-CNN), 1.35% (C-2D-CNN), 2.57% (3D-CNN), 0.34% (SpecAttenNet), 0.39% (CAPSNET), 0.33% (DPRESNET), 0.38% (SSRN), 2.42% (DRNN), and 0.12% (2DOC-SSAN). The kappa coefficient's increases are 16.04% (SVM), 13.42% (RF-200), 4.41% (Conv-Deconv-Net), 3.48% (2D-CNN), 3.36% (C-2D-CNN), 6.62% (3D-CNN), 0.84% (SpecAttenNet), 0.12% (CAPSNET), 0.33% (DPRESNET), 0.29% (SSRN), 2.41% (DRNN), and 0.40% (2DOC-SSAN). These promising results illustrate that both the spectral information and spatial information are fully captured by the proposed 3D-OCM, SSAM, and SSICM models.

Furthermore, by observing the different categories, it is obvious that our method outperforms the other counterparts in most cases. An encouraging observation is that the 3DOC-SSAN model can get superior results in some categories that are hard to identify, such as "Corn-mintill." For this land cover, the highest performance among all of the compared methods is obtained by 2D-CNN (99.40%). However, our model can achieve performance as high as 99.47%. There is another point that we want to touch on, i.e., the comparison between 2DOC-SSAN and 3DOC-SSAN. The only difference between these two models is that the convolution operation used in the Oct-Conv block. Compared with the 2-D convolution that focuses on exploring the information from the single feature map, 3-D convolution can mine the rich knowledge from all of the feature maps at the same time. This characteristic makes 3-D convolution more suitable to extract the features from the HSIs, which is proved by the classification results. The encouraging results discussed earlier demonstrate that our 3DOC-SSAN network is effective for the Indian Pines data set.

2) Analysis of University of Pavia Data Set: The visual and numerical classification results of different methods counted on the University of Pavia data set are shown in Fig. 11 and Table VII. As shown in Fig. 11, the classification map obtained by our method is close to the ground-truth map. Almost all of the samples can be predicted correctly and the boundaries of different categories are clear. As shown in Table VII, the behavior of our model is the strongest. The OA, AA, and Kappa scores are 99.87%, 99.85%, and 99.82%, respectively. Compared with other methods, the increases of OA scores obtained by our model are 7.41% (SVM), 4.26% (RF-200), 1.06% (Conv-Deconv-Net), 0.45% (2D-CNN), 0.19% (C-2D-CNN), 0.21% (3D-CNN), 0.12% (SpecAttenNet), 0.20% (CAPSNET), 0.72% (DPRESNET),

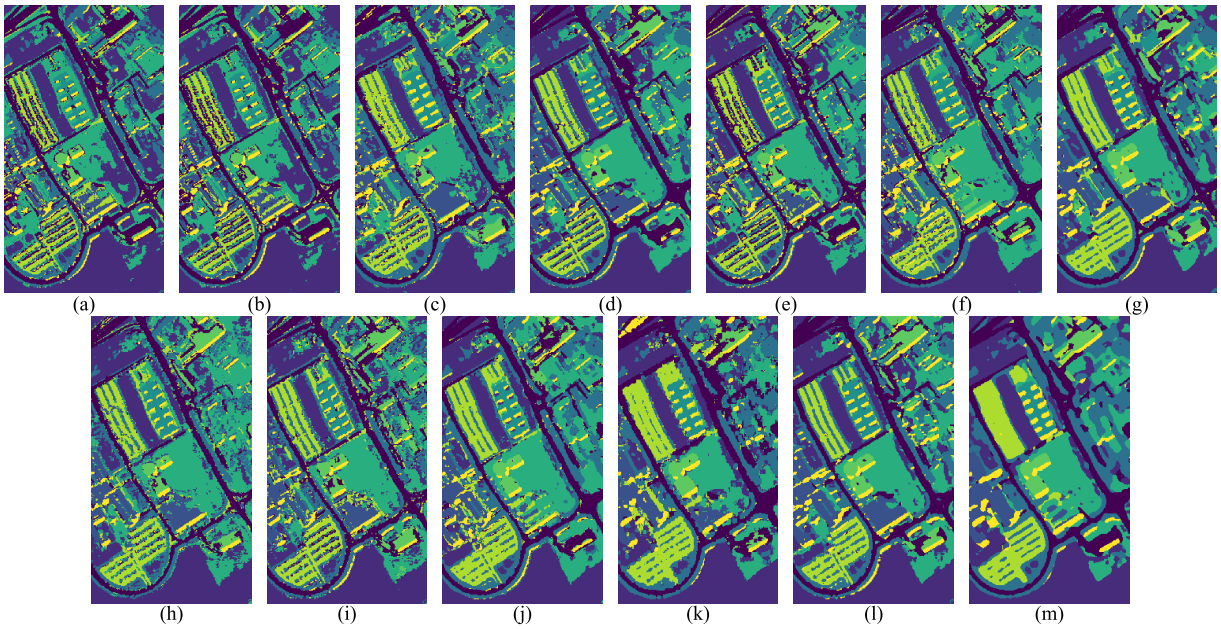


Fig. 11. Classification maps of different methods on the University of Pavia data set. (a) SVM (92.46%). (b) RF-200 (95.61%). (c) Conv-Deconv-Net (98.81%). (d) 2D-CNN (99.42%). (e) C-2D-CNN (99.68%). (f) 3D-CNN (99.66%). (g) SpecAttenNet (99.75%). (h) CAPSNET (99.67%). (i) DPRESNET (99.15%). (j) SSRN (99.78%). (k) DRNN (99.44%). (l) 2DOC-SSAN (99.82%). (m) 3DOC-SSAN (99.87%).

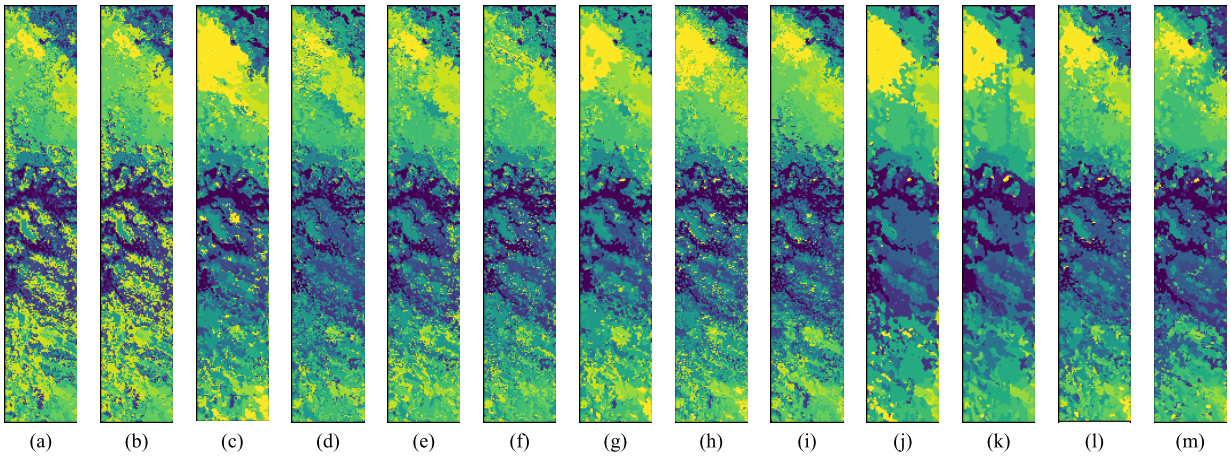


Fig. 12. Classification maps of different methods on the Botswana data set. (a) SVM (95.22%). (b) RF-200 (95.35%). (c) Conv-Deconv-Net (98.27%). (d) 2D-CNN (98.34%). (e) C-2D-CNN (98.53%). (f) 3D-CNN (97.18%). (g) SpecAttenNet (98.97%). (h) CAPSNET (98.45%). (i) DPRESNET (97.26%). (j) SSRN (99.21%). (k) DRNN (99.01%). (l) 2DOC-SSAN (99.34%). (m) 3DOC-SSAN (99.66%).

0.09% (SSRN), 0.43% (DRNN), and 0.05% (2DOC-SSAN). The improvements of AA scores are 6.28% (SVM), 2.79% (RF-200), 0.92% (Conv-Deconv-Net), 0.48% (2D-CNN), 0.23% (C-2D-CNN), 0.28% (3D-CNN), 0.08% (SpecAttenNet), 0.15% (CAPSNET), 1.55% (DPRESNET), 0.20% (SSRN), and 0.35% (DRNN). Different from other comparison methods, the AA scores of 2DOC-SSAN and 3DOC-SSAN are same to each other. The enhancements of kappa coefficient are 9.9% (SVM), 5.76%(RF-200), 1.44% (Conv-Deconv-Net), 0.48% (2D-CNN), 0.24% (C-2D-CNN), 0.28% (3D-CNN), 0.17% (SpecAttenNet), 0.10% (CAPSNET), 0.95% (DPRESNET), 0.12% (SSRN), 0.43% (DRNN), and 0.01% (2DOC-SSAN). These experimental results illustrate that our network can capture more discriminative features.

For some categories in the University of Pavia data set, such as "Painted Metal Sheets" and "Bare Soi," our method can reach 100% classification accuracy. For other categories, the classification accuracies obtained by our method can also reach a high level. In addition, for the class "Gravel," the classification accuracies of networks with attention mechanism, i.e., SpecAttenNet, 2DOC-SSAN, and 3DOC-SSAN, are more than 99%, which are significantly better than the networks without attentional mechanisms. This successfully demonstrates that the attentional mechanism plays a positive role in feature learning. From the abovementioned discussion, it can be seen that our method is effective in the University of Pavia data set.

3) Analysis of Botswana Data Set: The visual and numerical results of different methods counted on the Botswana data set are shown in Fig. 12 and Table VIII. As shown in Fig. 12, we can find that our method can generate clear classification map. As shown in Table VIII, the values of OA, AA, and Kappa of our method are higher than that of other comparison methods. The increases of OA

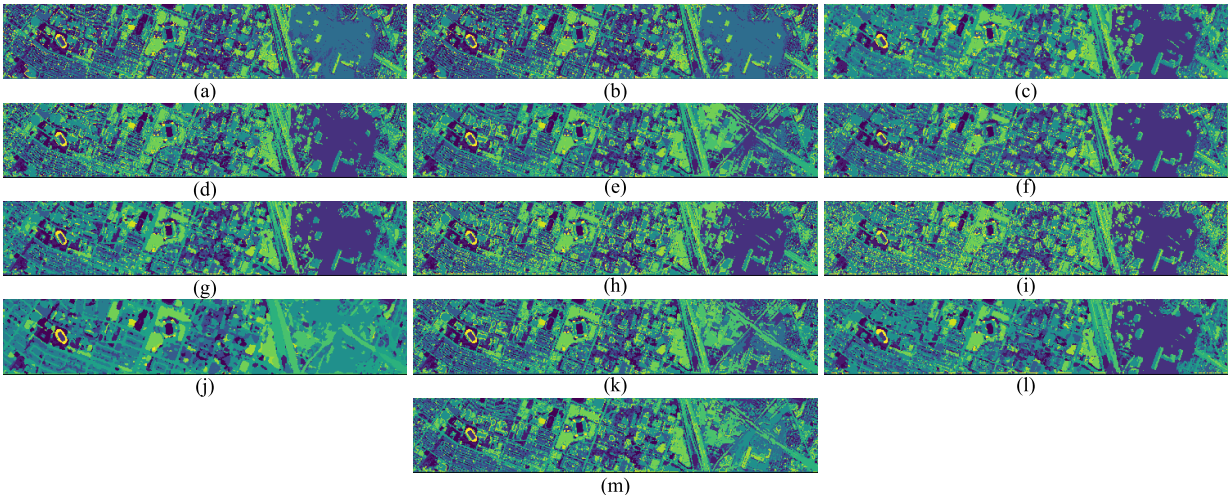


Fig. 13. Classification maps of different methods on the Houston data set. (a) SVM (66.91%). (b) RF-200 (74.43%). (c) Conv-Deconv-Net (78.57%). (d) 2D-CNN (80.50%). (e) C-2D-CNN (85.27%). (f) 3D-CNN (79.73%). (g) SpecAttenNet (79.51%). (h) CAPSNET (85.96%). (i) DPRESNET (81.62%). (j) SSRN (86.12%). (k) DRNN (85.81%). (l) 2DOC-SSAN (85.26%). (m) 3DOC-SSAN (87.59%).

scores obtained by our model are 4.44% (SVM), 4.31% (RF-200), 1.39% (Conv-Deconv-Net), 1.32% (2D-CNN), 1.13% (C-2D-CNN), 2.48% (3D-CNN), 0.69% (SpecAttenNet), 1.21% (CAPSNET), 2.40% (DPRESNET), 0.45% (SSRN), 0.65% (DRNN), and 0.32% (2DOC-SSAN). The improvements of AA scores are 5.66% (SVM), 3.42% (RF-200), 1.45% (Conv-Deconv-Net), 1.76% (2D-CNN), 0.99% (C-2D-CNN), 2.22% (3D-CNN), 0.69% (SpecAttenNet), 1.39% (CAPSNET), 2.15% (DPRESNET), 0.40% (SSRN), 0.75% (DRNN), and 0.32% (2DOC-SSAN). The enhancements of kappa coefficient are 5.21% (SVM), 4.59% (RF-200), 1.51% (Conv-Deconv-Net), 1.43% (2D-CNN), 1.23% (C-2D-CNN), 2.69% (3D-CNN), 0.75% (SpecAttenNet), 1.24% (CAPSNET), 2.61% (DPRESNET), 0.65% (SSRN), 0.71% (DRNN), and 0.34% (2DOC-SSAN).

It is encouraging that the proposed 3DOC-SSAN achieves the best performance in most of the categories. For nine categories, including "Water," "Hippo grass," "Floodplain grasses1," "Firescar2," "Island interior," "Acacia shrublands," "Short mopane," "Mixed mopane," and "Exposed soils," our method can obtain 100% classification accuracy. For the other five categories, the classification accuracy of our model exceeds 98.85%. Since the volume of labeled samples is small, but the size of Botswana data is large, the distribution of labeled samples of some categories (such as "Reeds1") is relatively scattered. In other words, there are many interference pixels around the center pixel, which would influence the classification results negatively [56]. Fortunately, our method still works on these categories. Taking "Reeds1" as an example, the highest performance among all of the comparisons is 97.69% (DRNN). However, our model can achieve performance as high as 98.86%. The above-discussed results demonstrate that our 3DOC-SSAN network is also effective for the Botswana data set.

4) Analysis of Houston Data Set: To study the performance of 3DOC-SSAN on the Houston HSI data set, we use the training set to train our model under the experimental settings mentioned in Section IV-B. Then, the trained network is utilized to predict the testing data. The visual and numerical results of different methods counted on the Houston data set are shown in Fig. 13 and Table IX. As shown in Fig. 13, the classification map acquired by 3DOC-SSAN is close to the original image. For the fuzzy parts of the original image, we can also predict most of the categories well. As shown in Table IX, the increases of the OA score obtained by 3DOC-SSAN are 20.68% (SVM), 13.16% (RF-200), 9.02% (Conv-Deconv-Net), 7.09% (2D-CNN), 2.32% (C-2D-CNN), 7.86% (3D-CNN), 8.08% (SpecAttenNet), 1.63% (CAPSNET), 5.97% (DPRESNET), 1.47% (SSRN), 1.78% (DRNN), and 2.33% (2DOC-SSAN). The improvements of the AA score achieved by our method are 22.11% (SVM), 14.98% (RF-200), 10.97% (Conv-Deconv-Net), 8.37% (2D-CNN), 3.33% (C-2D-CNN), 11.28% (3D-CNN), 9.41% (SpecAttenNet), 2.70% (CAPSNET), 7.30% (DPRESNET), 2.50% (SSRN), 2.25% (DRNN), and 3.03% (2DOC-SSAN). The enhancements of the Kappa coefficient obtained by 3DOC-SSAN are 22.27% (SVM), 13.18% (RF-200), 9.45% (Conv-Deconv-Net), 7.58% (2D-CNN), 3.45% (C-2D-CNN), 8.87% (3D-CNN), 8.65% (SpecAttenNet), 2.71% (CAPSNET), 6.10% (DPRESNET), 1.37% (SSRN), 1.83% (DRNN), and 1.84% (2DOC-SSAN).

For most categories, our method achieves good performance. For example, the accuracy of 3DOC-SSAN can be reached 100% on "Soil" and "Tennis Cou." However, since the training and testing sets within this HSI are defined apart (which increases the difficulty of classification), the results of different methods on some categories are not satisfactory, such as "Commercial" and "Highway." Even though, our method still obtains the best performance among all of the methods. Taking "Commercial" as an example, the highest performance among all of the compared methods is 77.82% (SSRN). Nevertheless, our model can achieve performance as high as 79.54%. The above-discussed encouraging results

TABLE VI

CLASSIFICATION RESULTS (%) OF DIFFERENT METHODS ON THE INDIAN PINES DATA SET. THE NAMES OF DIFFERENT CLASSES CAN BE FOUND IN TABLE I

CLASS	SVM	RF-200	Conv-Deconv-Net	2D-CNN	C-2D-CNN	3D-CNN	Spec-Atten-Net	CAPS-NET	DPRES-NET	SSRN	DRNN	2DOC-SSAN	3DOC-SSAN
1	100.00 ±0.00	100.00 ±0.00	100.00 ±0.00	93.75 ±0.74	100.00 ±0.00	98.75 ±0.56	100.00 ±0.00	97.34 ±0.42	98.75 ±0.45	97.98 ±0.38	96.14 ±0.53	**100.00** ± 0.00	**100.00** ± 0.00
2	69.92 ±0.57	75.85 ±0.62	92.35 ±1.54	94.60 ±0.81	95.46 ±0.36	89.32 ±0.71	97.28 ± 0.26	98.63± 0.26	98.38 ± 0.25	**98.72** ±0.19	95.70 ± 0.20	97.18 ± 0.29	**98.72** ±0.16
3	88.65 ±2.51	92.31 ±2.34	96.97 ±2.49	99.40 ±0.46	98.62 ±0.89	95.18 ±1.34	99.29 ±0.60	99.02 ±0.52	99.28 ±0.46	99.21 ±0.41	97.93 ±0.83	99.38 ±0.40	**99.47** ±0.40
4	97.81 ±0.91	97.37 ±0.83	99.69 ±1.34	99.69 ±0.29	**99.85** ±0.11	99.71 ±0.19	99.78 ±0.17	99.11 ±0.23	99.33 ±0.39	99.64 ±0.30	97.75 ±1.42	**99.85** ±0.14	**99.85** ±0.14
5	94.83 ±2.97	97.71 ±1.36	98.38 ±1.20	99.70 ±0.24	**99.76** ±0.20	99.28 ±0.51	99.70 ±0.21	98.59 ±0.98	97.79 ±0.82	98.71 ±0.39	98.71 ± 0.66	99.70 ±0.27	**99.76** ±0.23
6	98.79 ±0.31	99.35 ±0.24	98.83 ±0.72	99.65 ±0.32	99.35 ±0.45	98.93 ±0.86	99.65 ±0.25	97.79 ±0.76	99.75 ±0.16	98.89 ±0.93	98.65 ±0.56	**99.76** ±0.22	**99.76** ±0.23
7	100.00 ±0.00	95.00 ±2.76	100.00 ±0.00	98.93 ±0.27	100.00 ±0.00	100.00 ±0.00	100.00 ±0.00	100.00 ±0.00	100.00 ±0.00	100.00 ±0.00	100.00 ±0.00	**100.00** ±0.00	**100.00** ±0.00
8	100.00 ±0.00	100.00 ±0.00	100.00 ±0.00	99.89 ±0.15	100.00 ±0.00	100.00 ±0.00	100.00 ±0.00	100.00 ±0.00	100.00 ±0.00	100.00 ±0.00	100.00 ±0.00	**100.00** ±0.00	**100.00** ±0.00
9	100.00 ±0.00	96.00 ±2.46	100.00 ±0.00	100.00 ±0.00	100.00 ±0.00	100.00 ±0.00	100.00 ±0.00	100.00 ±0.00	96.66 ±2.36	100.00 ± 0.00	98.33 ±1.97	**100.00** ±0.00	**100.00** ±0.00
10	84.87 ±5.47	92.99 ±4.76	93.29 ±2.57	95.26 ±2.34	97.81 ±1.29	93.53 ±2.76	98.56 ±1.63	97.89 ± 1.24	98.77 ±1.12	98.62 ±1.25	97.45 ±1.39	98.66 ±1.24	**98.79** ±1.02
11	78.75 ±2.05	80.00 ± 1.96	92.60 ±1.34	95.53 ±1.34	92.32 ±1.21	88.30 ±2.43	97.54 ±1.21	**98.91** ±1.04	98.86 ±0.98	98.85 ±0.87	95.48 ±1.14	98.45 ±0.79	**98.91** ±0.66
12	90.98 ±6.47	93.95 ±3.76	96.69 ±2.93	98.32 ±1.47	98.28 ±1.36	98.67 ±1.05	98.78 ±1.14	98.02 ±1.36	97.68 ±1.82	**98.81** ±0.67	96.73 ±1.63	98.60 ±1.29	**98.81** ±1.08
13	100.00 ±0.00	99.27 ±0.82	100.00 ±0.00	100.00 ±0.00	100.00 ±0.00	100.00 ±0.00	100.00 ±0.00	99.87 ±0.12	100.00 ±0.00	100.00 ±0.00	99.39 ±0.24	**100.00** ±0.00	**100.00** ±0.00
14	96.56 ±0.96	97.56 ±0.85	98.69 ±0.24	99.19 ±0.36	97.55 ±0.47	98.17 ±0.51	99.21 ±0.62	99.20 ±0.22	99.41 ±0.39	99.17 ±0.73	98.95 ±0.21	**99.96** ±0.12	99.63 ±0.33
15	79.76 ±3.91	66.36 ±4.25	97.14 ±1.07	92.86 ±2.34	92.09 ±1.58	92.62 ±3.62	98.39 ±1.53	99.16 ±0.73	98.94 ± 0.65	98.89 ±0.71	97.64 ±1.36	**99.17** ±0.35	98.89 ±1.02
16	100.00 ±0.00	100.00 ±0.00	99.07 ±0.65	100.00 ±0.00	100.00 ±0.00	99.07 ±0.68	99.07 ±0.73	98.21 ±0.38	97.29 ±1.24	98.76 ±0.47	98.93 ±0.81	**100.00** ±0.00	**100.00** ±0.00
OA	85.25 ±0.59	87.48 ±0.67	95.34 ±0.49	97.02 ±0.43	96.24 ±0.38	93.42 ±0.55	98.50 ±0.26	99.01 ±0.31	98.99 ±0.32	98.98 ±0.21	96.86 ±0.43	98.89 ±0.19	**99.14** ±0.16
AA	92.56 ± 0.98	92.73 ±1.16	97.73 ±0.85	97.92 ±0.74	98.19 ±0.68	96.97 ±1.02	99.20 ±0.54	99.15 ±0.58	99.21 ±0.63	99.16 ±0.42	97.12 ±0.78	99.42 ±0.36	**99.54** ±0.29
K	82.96 ±0.87	85.58 ±1.04	94.59 ±0.76	96.52 ±0.68	95.64 ±0.65	92.38 ±0.97	98.16 ±0.49	98.88 ±0.55	98.67 ±0.59	98.71 ±0.47	96.59 ±0.62	98.60 ±0.29	**99.00** ±0.26

TABLE VII

CLASSIFICATION RESULTS (%) OF DIFFERENT METHODS ON THE UNIVERSITY OF PAVIA DATA SET. THE NAMES OF DIFFERENT CLASSES CAN BE FOUND IN TABLE II

CLASS	SVM	RF-200	Conv-Deconv-Net	2D-CNN	C-2D-CNN	3D-CNN	Spec-Atten-Net	CAPS-NET	DPRES-NET	SSRN	DRNN	2DOC-SSAN	3DOC-SSAN
1	89.23 ±1.74	96.04 ±1.28	97.92 ±0.69	99.34 ±0.43	99.31 ±0.47	99.27 ±0.36	99.66 ±0.25	99.67 ±0.21	98.92 ±0.34	99.77 ±0.31	97.82 ±0.19	99.70 ±0.23	**99.82** ±0.14
2	92.87 ±0.23	93.79 ±0.26	99.02 ±0.20	99.56 ±0.18	99.85 ±0.12	99.84 ±0.12	99.76 ±0.19	99.59 ±0.24	99.84 ±0.17	99.78 ±0.10	98.91 ±0.07	99.84 ±0.08	**99.93** ±0.05
3	83.57 ±1.36	92.42 ±1.75	98.28 ±1.23	98.83 ±0.97	98.71 ±0.86	99.28 ±0.98	99.69 ±0.41	99.32 ±0.45	97.38 ±1.48	98.34 ±0.47	97.51 ±0.22	**99.73** ±0.20	99.58 ±0.36
4	98.46 ±0.34	99.29 ±0.45	99.20 ±0.52	98.97 ±0.81	99.71 ±0.29	99.75 ±0.22	99.78 ±0.20	99.71 ±0.26	99.09 ±0.62	99.65 ±0.27	98.80 ±0.44	99.74 ±0.24	**99.81** ±0.19
5	99.96 ±0.12	99.93 ±0.14	100.00 ±0.00	99.64 ±0.27	100.00 ±0.00	100.00 ±0.00	99.82 ±0.16	99.84 ±0.15	99.87 ±0.12	99.64 ±0.21	99.80 ±0.19	99.90 ±0.06	**100.00** ±0.00
6	90.16 ±0.58	96.62 ±0.47	98.72 ±0.94	99.82 ±0.23	99.84 ±0.17	99.87 ±0.10	99.98 ±0.02	100.00 ±0.00	99.77 ±0.25	100.00 ±0.00	99.95 ± 0.02	99.97 ±0.03	**100.00** ±0.00
7	92.73 ±1.24	97.28 ±0.98	98.83 ±1.12	99.15 ±0.63	99.54 ±0.44	98.79 ±0.83	99.75 ±0.16	99.52 ±0.28	92.00 ±3.47	99.76 ±0.19	99.59 ±0.13	**99.85** ±0.10	99.76 ±0.21
8	95.52 ±1.36	98.21 ±0.62	98.86 ±0.83	99.05 ±0.75	99.60 ±0.31	99.36 ±0.53	99.54 ±0.29	99.73 ±0.14	98.57 ±0.89	99.12 ±0.32	98.64 ±0.37	**99.95** ±0.03	99.80 ±0.16
9	99.61 ±0.27	100.00 ±0.00	99.58 ±0.13	100.00 ±0.00	100.00 ±0.00	99.97 ±0.02	100.00 ±0.00	100.00 ±0.00	99.26 ±0.36	99.82 ±0.06	97.87 ±0.94	**100.00** ±0.00	**100.00** ±0.00
OA	92.46 ±0.94	95.61 ±0.87	98.81 ±0.42	99.42 ±0.39	99.68 ±0.24	99.66 ±0.30	99.75 ±0.17	99.67 ±0.21	99.15 ±0.38	99.78 ±0.12	99.44 ±0.25	99.82 ±0.14	**99.87** ±0.08
AA	93.57 ±1.35	97.06 ±0.68	98.93 ±0.67	99.37 ±0.54	99.62 ±0.26	99.57 ±0.23	99.77 ±0.20	99.70 ±0.21	98.30 ±0.82	99.65 ±0.18	99.50 ±0.24	**99.85** ±0.18	**99.85** ±0.15
K	89.92 ±1.27	94.06 ±0.95	98.38 ±0.53	99.34 ±0.44	99.58 ±0.39	99.54 ±0.42	99.65 ±0.28	99.64 ±0.27	98.87 ±0.67	99.70 ±0.25	99.39 ±0.26	99.81 ±0.16	**99.82** ±0.16

TABLE VIII

CLASSIFICATION RESULTS (%) OF DIFFERENT METHODS ON THE BOTSWANA DATA SET. THE NAMES OF DIFFERENT CLASSES CAN BE FOUND IN TABLE III

CLASS	SVM	RF-200	Conv-Deconv-Net	2D-CNN	C-2D-CNN	3D-CNN	Spec-Atten-Net	CAPS-NET	DPRES-NET	SSRN	DRNN	2DOC-SSAN	3DOC-SSAN
1	99.50 ±0.32	100.00 ±0.00	99.50 ±0.36	98.79 ±0.89	100.00 ±0.00	99.92 ±0.12	99.83 ±0.16	98.86 ±0.78	99.79 ±0.18	98.31 ±0.94	97.71 ±1.25	**100.00** ±0.00	**100.00** ±0.00
2	98.87 ±1.04	100.00 ±0.00	99.15 ±0.75	98.91 ±0.56	100.00 ±0.00	99.15 ±0.43	98.59 ±0.39	98.59 ±0.58	98.59 ±0.51	99.12 ±0.21	99.29 ±0.33	**100.00** ±0.00	**100.00** ±0.00
3	96.29 ±1.40	96.20 ±1.59	99.82 ±0.15	99.04 ±0.82	100.00 ±0.00	97.20 ±1.29	99.55 ±0.38	98.92 ±0.54	99.09 ±0.73	99.24 ±0.51	99.66 ±0.26	**100.00** ±0.00	**100.00** ±0.00
4	79.78 ±3.47	97.84 ±1.45	99.34 ±0.51	99.24 ±0.73	99.56 ±0.44	99.46 ±0.62	99.62 ±0.06	99.34 ±0.42	98.54 ±1.09	98.68 ±0.37	99.65 ±0.25	**99.67** ±0.16	99.60 ±0.36
5	78.42 ±4.69	84.77 ±3.95	90.79 ±1.37	94.27 ±1.25	91.05 ±1.93	85.52 ±3.67	93.30 ±1.29	96.84 ±1.08	92.05 ±2.46	96.54 ±1.78	97.69 ±1.93	96.40 ±2.41	**98.86** ±1.05
6	90.63 ±0.34	86.94 ±1.39	96.40 ±1.07	91.64 ±0.55	94.39 ±0.37	93.55 ±0.86	96.65 ±0.38	97.66 ± 0.45	93.82 ±1.02	98.75 ±0.28	98.75 ±0.49	97.24 ±0.63	**98.87** ±0.21
7	100.00 ±0.00	99.65 ±0.34	99.74 ±0.25	99.35 ±0.41	100.00 ±0.00	99.83 ±0.16	100.00 ±0.00	98.60 ±0.27	98.36 ±0.31	99.43 ±0.56	100.00 ±0.00	**100.00** ±0.00	**100.00** ±0.00
8	95.03 ±2.92	99.77 ±0.36	100.00 ±0.00	98.87 ±1.04	100.00 ±0.00	99.77 ±0.24	99.88 ±0.06	97.98 ±0.83	97.83 ±0.79	98.86 ±0.17	98.84 ±0.68	**100.00** ±0.00	**100.00** ±0.00
9	90.85 ±1.45	86.72 ±3.94	98.52 ±1.16	98.24 ±1.07	97.89 ±0.62	97.54 ±1.33	99.65 ±0.43	98.38 ±0.92	93.22 ±1.82	99.77 ±0.17	99.81 ±0.29	99.79 ±0.16	**99.83** ±0.05
10	100.00 ±0.00	98.99 ±0.34	100.00 ±0.00	98.62 ±0.59	100.00 ±0.00	100.00 ±0.00	99.91 ±0.07	99.67 ±0.13	99.35 ±0.41	98.76 ±0.35	100.00 ± 0.00	99.17 ±0.32	**100.00** ±0.00
11	98.04 ±0.49	99.13 ±0.56	99.49 ±0.24	97.45 ±1.34	**99.78** ±0.15	97.16 ±0.98	99.49 ±0.09	97.31 ±0.72	98.82 ±0.92	98.87 ±0.46	98.00 ±0.5	**99.78** ±0.13	99.36 ±0.41
12	96.16 ±1.24	100.00 ±0.00	95.49 ±0.87	98.89 ±0.89	100.00 ±0.00	97.02 ±1.48	100.00 ±0.00	98.43 ±0.64	98.51 ±0.69	97.68 ±1.02	97.52 ±1.59	**100.00** ±0.00	**100.00** ±0.00
13	93.70 ±0.52	98.57 ±0.92	98.81 ±1.04	98.74 ±0.76	100.00 ±0.00	99.58 ±0.37	100.00 ±0.00	99.01 ±0.20	98.43 ±0.27	99.39 ±0.36	99.58 ±0.25	**100.00** ±0.00	**100.00** ±0.00
14	100.00 ±0.00	100.00 ±0.00	99.08 ±0.75	99.83 ±0.12	100.00 ±0.00	99.69 ±0.04	100.00 ±0.00	100.00 ±0.00	99.65 ±0.22	100.00 ±0.00	100.00 ±0.00	**100.00** ±0.00	**100.00** ±0.00
OA	95.22 ±1.09	95.35 ±0.94	98.27 ±0.71	98.34 ±0.69	98.53 ±0.35	97.18 ±0.41	98.97 ±0.36	98.45 ±0.48	97.26 ±0.84	99.21 ±0.52	99.01 ±0.45	99.34 ±0.32	**99.66** ±0.19
AA	94.09 ±1.58	96.33 ±1.23	98.30 ±0.92	97.99 ±0.85	98.76 ±0.56	97.53 ±0.78	99.06 ±0.44	98.36 ±0.81	97.60 ±1.02	99.35 ±0.63	99.00 ±0.51	99.43 ±0.40	**99.75** ±0.28
K	94.42 ±1.49	95.04 ±1.36	98.12 ±0.82	98.20 ±0.84	98.40 ±0.63	96.94 ±0.67	98.88 ±0.58	98.39 ±0.71	97.02 ±0.95	98.98 ±0.39	98.92 ±0.47	99.29 ±0.36	**99.63** ±0.33

TABLE IX

CLASSIFICATION RESULTS (%) OF DIFFERENT METHODS ON THE HOUSTON DATA SET. THE NAMES OF DIFFERENT CLASSES CAN BE FOUND IN TABLE IV

CLASS	SVM	RF-200	Conv-Deconv-Net	2D-CNN	C-2D-CNN	3D-CNN	Spec-Atten-Net	CAPS-NET	DPRES-NET	SSRN	DRNN	2DOC-SSAN	3DOC-SSAN
1	80.53 ±3.89	80.98 ±2.45	78.79 ±3.71	82.52 ±1.76	82.52 ±1.02	80.24 ± 2.54	81.86 ±1.32	82.09 ± 1.22	80.59 ±2.06	81.57 ±1.13	82.17 ± 2.01	82.63 ± 1.21	**82.91** ± 1.04
2	76.22 ± 1.24	79.61 ± 1.29	80.64 ±1.51	83.08 ± 1.07	**85.15** ± 0.86	74.81 ± 1.46	81.01 ± 0.72	**85.15** ± 0.67	82.67 ± 0.79	84.49 ± 0.52	85.06 ± 1.31	84.58 ± 0.84	**85.15** ± 0.50
3	50.69 ±3.91	55.45 ± 5.78	63.34 ± 3.82	64.75 ± 3.29	90.09 ± 1.24	58.01 ± 2.86	61.38 ± 2.67	94.65 ± 0.96	79.52 ± 2.73	96.23 ± 0.82	95.04 ± 1.26	93.37 ± 1.31	**96.41** ± 0.69
4	82.67 ± 1.29	85.05 ± 0.92	89.40 ± 0.81	79.73 ± 0.84	92.99 ± 0.65	91.75 ± 0.58	89.67 ± 0.55	92.89 ±0.40	86.08 ± 0.79	89.96 ± 0.35	92.63 ± 0.57	88.45 ± 0.41	**93.03** ± 0.32
5	93.75 ± 1.74	97.62 ± 0.81	96.49 ± 0.92	99.71 ± 0.21	100.00 ± 0.00	92.61 ± 1.87	96.21 ± 0.95	99.91 ± 0.04	97.23 ± 0.48	100.00± 0.00	100.00 ± 0.00	**100.00** ± 0.00	**100.00** ± 0.00
6	69.23 ± 3.47	80.50 ± 1.27	81.12 ± 1.14	93.06 ± 0.52	95.21 ± 0.31	93.00 ± 0.49	84.92 ± 1.03	94.16 ± 0.41	93.14 ± 0.58	95.18 ± 0.26	94.35 ± 0.46	89.03 ± 0.98	**95.58** ± 0.21
7	80.95 ± 0.46	76.66 ± 0.87	85.36 ± 0.57	83.02 ± 0.39	82.28 ± 0.36	**85.64** ± 0.31	81.25 ± 0.89	83.30 ± 1.04	80.97 ± 0.85	82.27 ± 0.62	83.11 ± 0.47	81.91 ± 0.53	85.59 ± 0.46
8	37.13 ± 8.93	44.26 ± 7.38	69.89 ± 3.92	74.45 ± 3.81	65.05 ± 1.49	55.56 ± 3.60	66.57 ± 1.63	62.77 ± 0.93	70.06 ± 0.89	77.82 ± 0.36	63.72 ± 1.27	76.79 ± 0.56	**79.54** ± 0.40
9	78.56 ± 0.91	80.72 ± 0.72	81.21 ± 0.69	83.57 ± 0.85	85.55 ± 0.55	84.71 ± 0.64	84.14 ± 0.60	79.61 ± 0.99	82.01 ± 0.38	84.31 ± 0.36	80.26 ± 0.80	81.30 ± 0.39	**85.85** ± 0.33
10	40.51 ± 1.85	39.34 ± 2.03	58.38 ± 1.47	56.08 ± 1.58	56.27 ± 0.81	43.14 ± 5.95	53.86 ± 4.87	65.91 ± 0.62	52.45 ± 2.81	50.48 ± 3.76	64.38 ± 2.49	59.83 ± 1.32	**66.20** ± 2.63
11	48.39 ± 3.73	65.18 ± 1.28	60.92 ± 1.03	72.58 ± 1.93	**91.36** ± 1.40	76.03 ± 2.89	70.88 ± 1.65	86.05 ± 0.39	74.93 ± 0.76	89.41 ± 0.45	91.35 ± 0.55	89.26 ± 1.69	89.72 ± 0.83
12	76.17 ± 2.99	76.57 ± 2.73	92.22 ± 1.04	91.64 ± 1.47	85.59 ± 1.93	81.75 ± 0.82	90.68 ± 0.37	93.76 ± 0.82	93.57 ± 0.36	94.75 ± 0.43	92.69 ± 1.21	93.08 ± 0.68	**95.01** ± 0.31
13	70.52 ± 1.64	85.71 ± 2.83	88.77 ± 0.58	84.91 ± 1.29	90.17 ± 0.37	90.52 ± 0.46	80.35 ± 1.25	88.77 ± 0.73	83.69 ± 0.84	83.85 ± 0.95	90.63 ± 0.60	85.61 ± 1.41	**90.75** ± 1.34
14	72.06 ± 2.34	89.50 ± 1.34	80.97 ± 1.84	89.87 ± 1.23	95.95 ± 0.86	79.76 ± 2.31	98.78 ± 0.45	97.57 ± 0.60	89.27 ± 1.05	100.00 ± 0.00	97.97 ± 0.44	**100.00** ± 0.00	**100.00** ± 0.00
15	57.24 ± 3.20	84.35 ± 1.53	74.21 ± 2.88	81.82 ± 1.55	96.09 ± 0.59	89.47 ± 1.07	83.51 ±1.92	96.19 ± 0.72	96.03 ± 0.89	95.73 ± 0.56	96.21 ± 0.47	94.87 ± 0.98	**96.53** ± 0.37
OA	66.91 ± 2.66	74.43 ± 1.34	78.57 ± 1.16	80.50 ±0.84	85.27 ± 0.74	79.73 ± 1.26	79.51 ± 1.08	85.96 ± 0.63	81.62 ± 0.97	86.12 ± 0.55	85.81 ± 0.72	85.26 ± 0.92	**87.59** ± 0.49
AA	67.64 ± 2.74	74.77 ± 1.43	78.78 ± 1.09	81.38 ± 0.92	86.42 ± 0.76	78.47 ± 1.33	80.34 ± 1.00	87.05 ± 0.75	82.45 ± 1.02	87.25 ± 0.54	87.50 ± 0.66	86.72 ± 0.83	**89.75** ± 0.52
K	64.22 ± 2.49	73.31 ± 1.27	77.04 ± 1.25	78.91 ± 0.77	83.04 ± 0.69	77.62 ± 1.40	77.84 ± 0.98	83.78 ± 0.59	80.39 ± 0.86	85.12 ± 0.63	84.66 ± 0.74	84.65 ± 0.85	**86.49** ± 0.55

TABLE X
OAs (%) WITH DIFFERENT FEATURE EXTRACTION MODELS ON THE FOUR DATA SETS

	2D-SSAN	3D-SSAN	2DOC-SSAN	3DOC-SSAN
Indian Pines	98.68±0.67	98.70±0.45	98.89±0.19	**99.14**±0.16
University of Pavia	99.70±0.21	99.75±0.12	99.82±0.14	**99.87**±0.08
Botswana	98.87±0.79	98.98±0.72	99.34±0.32	**99.66**±0.19
Houston	83.47±1.59	84.39±1.38	85.26±0.92	**87.59**±0.49

TABLE XI
OAs (%) WITH DIFFERENT ATTENTION MODELS ON THE FOUR DATA SETS

	3D-OCM	3D-OCM +SPAM	3D-OCM +SPEM	3D-OCM +SSAM
Indian Pines	98.13±0.68	98.35±0.45	98.39±0.46	**98.79**±0.23
University of Pavia	99.62±0.34	99.69±0.26	99.70±0.26	**99.76**±0.13
Botswana	98.34±0.85	98.69±0.64	98.75±0.59	**99.16**±0.42
Houston	80.91±2.32	83.87±1.84	83.79±1.68	**85.32**±0.87

demonstrate that our 3DOC-SSAN network is effective for the Houston data set.

D. Analysis of Submodels

1) Analysis of 3D-OCM: In order to verify the importance of 3D-OCM, we design three additional models to replace 3D-OCM for extracting features, i.e., a CNN based on 2-D convolution (CNN-2D), a CNN based on 3-D convolution (CNN-3D), and the octave convolution model based on 2-D convolution (2D-OCM). The classification networks based on four feature learning models are recorded 2D-SSAN, 3D-SSAN, 2DOC-SSAN, and 3DOC-SSAN here. Since there are four Oct-Conv blocks in 3D-OCM, we also use four convolution blocks in the other three models for fairness. For three test data sets, OA scores of different classification networks are summarized in Table X. It is obvious that 3DOC-SSAN outperforms the others. The reason behind this is that the 3D-OCM model can capture spatial and spectral information simultaneously. Compared with the other methods, the OA scores' enhancements obtained by our model are 0.46% (2D-SSAN), 0.44% (3D-SSAN), and 0.34% (2DOC-SSAN) on the Indian Pines data set, 0.17% (2D-SSAN), 0.12% (3D-SSAN), and 0.05% (2DOC-SSAN) on the University of Pavia data set, 0.79% (2D-SSAN), 0.68% (3D-SSAN), and 0.32% (2DOC-SSAN) on the Botswana data set, and 4.12% (2D-SSAN), 3.2% (3D-SSAN), and 2.33% (2DOC-SSAN). The above-discussed promising results demonstrate that the features obtained by 3D-OCM are comprehensive and representative.

2) Analysis of SSAM: In this section, we study the function of SSAM. As mentioned in Section III-B, SSAM is proposed to generate the attention maps from the spectral and spatial aspects. It consists of two components: the spatial attention model (SPAM) and spectral attention model (SPEM). To illustrate the importance of SSAM, we do the following experiments. First, we eliminate SSAM from our classification network, and we name it 3D-OCM for clear. Second, only the SPAM model is adopted in our network, and we record it 3D-OCM+SPAM for short. Third, only the SPEM model is added in our network, and we record it 3D-OCM+SPEM for convenience. Finally, the SSAM model is embedded in the classification network, and we name it 3D-OCM+SSAM. Since the input of the SSICM model requires two feature maps, but the first three comparative experiments only generate one feature map, we remove the SSICM model in all experiments. The OA scores of the four schemes counted on different data sets are shown in Table XI.

TABLE XII
OAs (%) WITH INFORMATION COMPLEMENTARITY ON THE FOUR DATA SETS

	3DOC-SSAN without SSICM	3DOC-SSAN
Indian Pines	98.79±0.23	**99.14**±0.16
University of Pavia	99.76±0.13	**99.87**±0.08
Botswana	99.16±0.42	**99.66**±0.19
Houston	85.32±0.87	**87.59**±0.49

It is obvious that the OA score of 3D-OCM + SSAM is the highest among the four schemes. On the Indian Pines data set, the OA score of 3D-OCM + SSAM is 98.79%, and it is 0.66%, 0.44%, and 0.40% higher than the OA values of 3D-OCM, 3D-OCM + SPAM, and 3D-OCM + SPEM. Similar to the Indian Pines data set, the increases of OA scores obtained by 3D-OCM + SSAM are 0.14% (3D-OCM), 0.07% (3D-OCM + SPAM), and 0.06% (3D-OCM + SPEM) on the University of Pavia data set, 0.82% (3D-OCM), 0.48% (3D-OCM + SPAM), and 0.41% (3D-OCM + SPEM) on the Botswana data set, and 4.41% (3D-OCM), 1.45% (3D-OCM + SPAM), and 1.53% (3D-OCM + SPEM) on the Houston data set. The results shown in Table IX show that the performance of the schemes with the attention mechanism (3D-OCM + SPAM, 3D-OCM + SPEM, and 3D-OCM + SSAM) is better than that of the scheme without attention method (3D-OCM). This demonstrates that the attention method plays a positive role in the HSIs' feature learning and classification. Moreover, since the SSAM model takes the spatial and spectral factors into account at the same time, 3D-OCM + SSAM performance is better than that of 3D-OCM + SPAM and 3D-OCM + SPEM. The experimental results show that the two attention models designed in this article can capture important spatial and spectral information and improve the discrimination of features.

3) Analysis of SSICM: SSICM is proposed to fuse the spatial and spectral features in a mutually complementary manner. Through SSICM, we can transmit important spatial information to spectral features and infuse significant spectral information to spatial features. In this section, we study this model by eliminating it from our classification network. In other words, we compare the classification results between 3DOC-SSAN and 3DOC-SSAN without SSICM. The OAs of the two experiments are shown in Table XII. From the observation of Table XII, we can easily find that the performance of 3DOC-SSAN is better than that of 3DOC-SSAN without SSICM in all scenarios. These results show that the information complementary model SSICM is useful to improve our network for the HSIs' classification task.

TABLE XIII

OAs (%) WITH DIFFERENT SUBMODELS ON THE FOUR DATA SETS

	NET_1	NET_2	NET_3
Indian Pines	98.13±0.68	98.79±0.23	**99.14**±0.16
University of Pavia	99.62±0.34	99.76±0.13	**99.87**±0.08
Botswana	98.34±0.85	99.16±0.42	**99.66**±0.19
Houston	80.91±2.32	85.32±0.87	**87.59**±0.49

TABLE XIV

NUMBER OF PARAMETERS ON DIFFERENT DATA SETS (M: MILLION)

	3D-OCM	SSAM	SSICM	3DOC-SSAN
Indian Pines	2.63	0.44	0	3.07
University of Pavia	1.35	0.12	0	1.47
Botswana	1.91	0.23	0	2.14
Houston	1.89	0.23	0	2.12

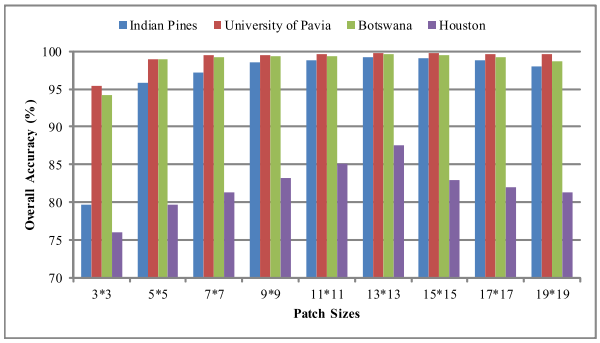


Fig. 14. OAs with different patch sizes on different data sets.

E. Ablation Study

In the 3DOC-SSAN model, there are three parts: 3D-OCM, SSAM, and SSICM. The 3D-OCM block aims to capture the spatial–spectral features from the HSIs. The SSAM block is introduced to improve the discrimination of the obtained features from the 3D-OCM block. The SSICM block is developed to integrate important information and remove redundant information. To study the contributions of each submodel to our method, we construct three networks to complete the HSI classification, as follows.

1) *NET_1:* 3D-OCM.
2) *NET_2:* 3D-OCM + SSAM.
3) *NET_3:* 3D-OCM + SSAM + SSICM.

Their OA values counted on four HSIs are summarized in Table XIII, where we can easily find that each component has its positive contributions to the classification task. Taking the Indian Pines data set as an example, we can find the following three points. First, the OA score of Net_1 is 98.13%. This classification accuracy is acceptable, which indicates that 3D-OCM is suitable for the HSI classification task. Second, by comparing the results of Net_1 and Net_2, we can find the attention mechanism is useful for improving the discrimination of features. Third, by observing the results of Net_2 and Net_3, it is apparent that the SSICM model can also bring an encouraging improvement (0.35%). The reason is that the important information within the outputs of different attention models is highlighted, and the redundant information is suppressed. These positive results demonstrate that each submodel can make positive contributions to the HSIs classification.

F. Impact of Patch Size

As mentioned in Section IV-B, we select the image patch centered at each pixel as the input of our 3DOC-SSAN, and the patch size is 13 × 13. However, the size of the patch has a great influence on the classification results [57], [58]. To obtain a suitable patch size for our model, we select image patches with different sizes for accomplishing the classification experiments. In detail, the image patches' sizes are varied from 3 × 3 to 19 × 19 with the interval of 2, and the OA values of our 3DOC-SSAN model counted on different HSI data sets are shown in Fig. 14. From the observation of the bars, we can find the weakest performance appears when the patch size equals 3 × 3. This illustrates that small patches bring less information, which cannot support our model to get a good performance. As the patch size becomes larger, each patch contains more spatial and spectral information, the performance of our method is increased to an appropriate extent, and the peak values for different data sets appear in 13 × 13. When the patch size is larger than 13 × 13, the performance of our method is decreased, especially for the Houston data set. The reason behind this is that, with the increase in the patch size, the redundant information contained in the patch will increase, which will harm the classification tasks. Therefore, when the size of the patch is too large or too small, the classification accuracy would be affected negatively. To sum up, we set the image patch size at 13 × 13 for our network.

G. Impact of Proportion of Training Samples

For the HSI classification, finding sufficient labeled samples to train a classifier is a difficult and time-consuming task. Therefore, the classification performance of a model with the limited training samples becomes an important assessment criterion. In the previous experiments, we fixed the number of samples to testify the performance of our model. The positive classification results counted on four data sets demonstrate the effectiveness of our 3DOC-SSAN model. In addition, to study if 3DOC-SSAN is useful or not when the number of training data is limited, and to observe the influence of various volumes of the training set on our method, we design the following experiments. For different data sets, we randomly select 1%, 3%, 5%, 10%, 15%, and 20% of samples to train our 3DOC-SSAN model. Then, the rest of the samples are used to test 3DOC-SSAN. Here, we adopt two networks, SSRN and DPRESNET, as the reference, which performs well in the small training set scenario.

The results counted on four data sets are summarized in Fig. 15, where we can find the following points. First, our 3DOC-SSAN outperforms the other two compared methods in most cases. Second, the performance of different networks is acceptable but satisfactory when there are only 1% samples in the training set. However, when the proportion of training samples increases, their performance is improved drastically. Especially, when the proportion of training samples changes from 1% to 10%, the behavior of various methods is enhanced dramatically. The above-discussed observations illustrate that: 1) the proposed 3DOC-SSAN can classify the HSIs well even

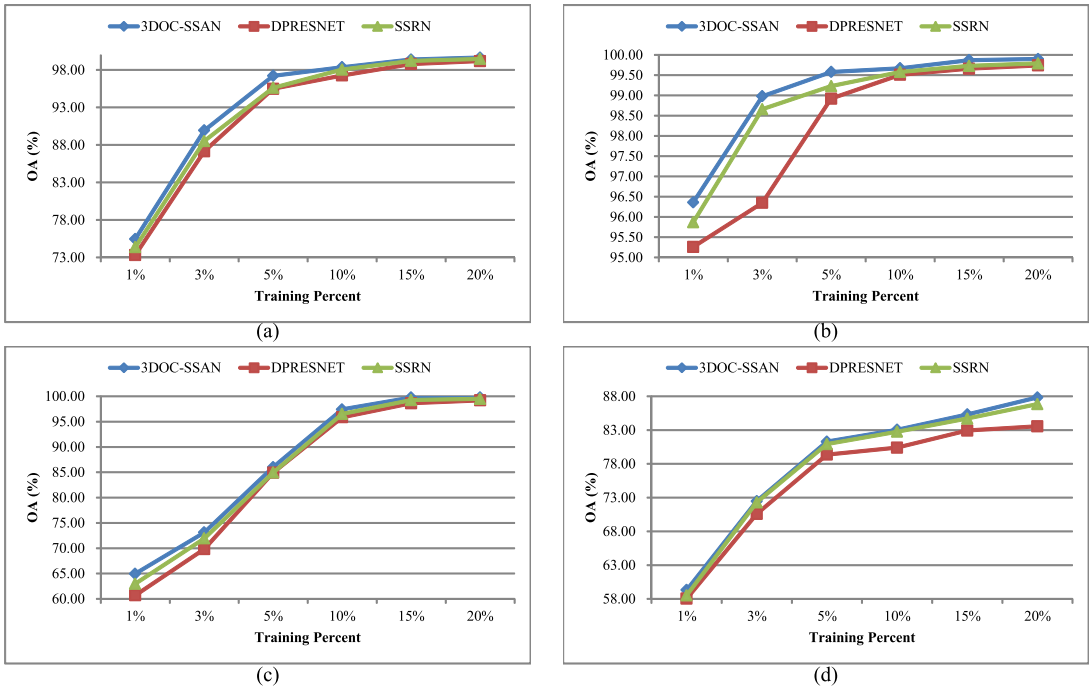


Fig. 15. OAs with different training percent of different data sets. (a) Indian Pines. (b) University of Pavia. (c) Botswana. (d) Houston.

though the number of the labeled data is few and 2) our 3DOC-SSAN outperform the other two popular methods no matter the number of the labeled data is enough or not.

H. Impact of the Number of Parameters

In DL, the network's complexity is an important assessment criterion that can be measured by the number of parameters apparently [59]. In this section, we study the number of parameters in our network. As mentioned in Section III, there are three submodels in 3DOC-SSAN: 3D-OCM, SSAM, and SSICM. Thus, we provide the numbers of parameters corresponding to each model for deeply studying the complexity of 3DOC-SSAN. The details are summarized in Table XIV, where the volume of parameters is counted on different data sets. We can find that the parameters of our model are concentrated upon the 3D-OCM submodel, which consists of an octave convolution network with the 3-D convolution. Since the main operations within the SSAM submodel are transposition and dot product, its parameters' volume is not too large. For the SSICM submodel, the number of parameters is zero as it only contains the reshaping and summation operations.

V. Conclusion

In this article, we have proposed an end-to-end network named 3DOC-SSAN for the HSI classification task, which can capture and highlight important spatial and spectral information. First, we use an octave convolution model that consists of four Oct-Conv blocks to process the spatial information for fusing high- and low-frequency information and reduce the number of network parameters. In order to process spatial and spectral information simultaneously, we have extended the octave convolution model to a 3-D version (named 3D-OCM), and then, the spatial–spectral features can be acquired simultaneously. Second, due to the characteristics of HSIs, two attention mechanisms from spatial and spectral aspects have been employed in our network. Through these models, the significant spatial areas and special spectral bands are highlighted to improve the discrimination of features. Finally, in order to ensure the integrity of the information, we have designed SSICM that can remain the important parts of different features. In SSICM, the information flows are established to transmit mutual information between spatial and spectral features. Not only important spatial information but also particular spectral information makes contributions to the classification tasks. Experiments with four common data sets of HSIs show that our method can obtain good results. However, since the use of 3-D convolution in our network, the training time of our method is relatively long. Therefore, our further work mainly focuses on decreasing the training time while ensuring the classification accuracy of HSI classification tasks.

References

[1] J. Feng, L. Jiao, F. Liu, T. Sun, and X. Zhang, "Unsupervised feature selection based on maximum information and minimum redundancy for hyperspectral images," *Pattern Recognit.*, vol. 51, pp. 295–309, Mar. 2016.

[2] Y. Xu, B. Du, F. Zhang, and L. Zhang, "Hyperspectral image classification via a random patches network," *ISPRS J. Photogramm. Remote Sens.*, vol. 142, pp. 344–357, Aug. 2018.

[3] X. Zhang, Y. Sun, K. Jiang, C. Li, L. Jiao, and H. Zhou, "Spatial sequential recurrent neural network for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 11, pp. 4141–4155, Nov. 2018.

[4] R. Heylen, M. Parente, and P. Gader, "A review of nonlinear hyperspectral unmixing methods," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 1844–1868, Jun. 2014.

[5] W. Li and Q. Du, "Collaborative representation for hyperspectral anomaly detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 3, pp. 1463–1474, Mar. 2015.

[6] X. Ma, X. Zhang, X. Tang, H. Zhou, and J. Licheng, "Hyperspectral anomaly detection based on low-rank representation with data-driven projection and dictionary construction," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 2226–2239, 2020.

[7] S. Delalieux, B. Somers, B. Haest, T. Spanhove, J. Vanden Borre, and C. A. Mücher, "Heathland conservation status mapping through integration of hyperspectral mixture analysis and decision tree classifiers," *Remote Sens. Environ.*, vol. 126, pp. 222–231, Nov. 2012.

[8] J. Ham, Y. Chen, M. M. Crawford, and J. Ghosh, "Investigation of the random forest framework for classification of hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 492–501, Mar. 2005.

[9] J. A. Gualtieri and S. Chettri, "Support vector machines for classification of hyperspectral data," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, vol. 2, Jul. 2000, pp. 813–815.

[10] J. Li, J. M. Bioucas-Dias, and A. Plaza, "Spectral–spatial classification of hyperspectral data using loopy belief propagation and active learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 2, pp. 844–856, Feb. 2013.

[11] W. Li and Q. Du, "A survey on representation-based classification and detection in hyperspectral remote sensing imagery," *Pattern Recognit. Lett.*, vol. 83, pp. 115–123, Nov. 2016.

[12] W. Li, C. Chen, H. Su, and Q. Du, "Local binary patterns and extreme learning machine for hyperspectral imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 7, pp. 3681–3693, Jul. 2015.

[13] A. Soltani-Farani, H. R. Rabiee, and S. A. Hosseini, "Spatial-aware dictionary learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 1, pp. 527–541, Jan. 2015.

[14] Y. Chen, N. M. Nasrabadi, and T. D. Tran, "Hyperspectral image classification using dictionary-based sparse representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 10, pp. 3973–3985, Oct. 2011.

[15] J. Zhao *et al.*, "Spectral–spatial classification of hyperspectral imagery with cooperative game," *ISPRS J. Photogramm. Remote Sens.*, vol. 135, pp. 31–42, Jan. 2018.

[16] J. Li, J. M. Bioucas-Dias, and A. Plaza, "Spectral–spatial hyperspectral image segmentation using subspace multinomial logistic regression and Markov random fields," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 3, pp. 809–823, Mar. 2012.

[17] Y. Chen, X. Zhao, and X. Jia, "Spectral–spatial classification of hyperspectral data based on deep belief network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 2381–2392, Jun. 2015.

[18] X. Yuan, B. Huang, Y. Wang, C. Yang, and W. Gui, "Deep learning-based feature representation and its application for soft sensor modeling with variable-wise weighted SAE," *IEEE Trans. Ind. Informat.*, vol. 14, no. 7, pp. 3235–3243, Jul. 2018.

[19] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1492–1500.

[20] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, "When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2811–2821, May 2018.

[21] Y. Xu, L. Zhang, B. Du, and F. Zhang, "Spectral–spatial unified networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 10, pp. 5893–5909, Oct. 2018.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[23] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 3856–3866.

[24] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[25] Y. Chen, J. Li, H. Xiao, X. Jin, S. Yan, and J. Feng, "Dual path networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 4467–4475.

[26] M. E. Paoletti, J. M. Haut, R. Fernandez-Beltran, J. Plaza, A. J. Plaza, and F. Pla, "Deep pyramidal residual networks for spectral–spatial hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 740–754, Feb. 2019.

[27] M. E. Paoletti, J. M. Haut, J. Plaza, and A. Plaza, "Deep&dense convolutional neural network for hyperspectral image classification," *Remote Sens.*, vol. 10, no. 9, p. 1454, 2018.

[28] M. E. Paoletti *et al.*, "Capsule networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 4, pp. 2145–2160, Apr. 2019.

[29] X. Kang, B. Zhuo, and P. Duan, "Dual-path network-based hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 3, pp. 447–451, Mar. 2019.

[30] M. E. Paoletti, J. M. Haut, J. Plaza, and A. Plaza, "A new deep convolutional neural network for fast hyperspectral image classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 120–147, Nov. 2018.

[31] Z. Meng, L. Li, X. Tang, Z. Feng, L. Jiao, and M. Liang, "Multipath residual network for spectral-spatial hyperspectral image classification," *Remote Sens.*, vol. 11, no. 16, p. 1896, Aug. 2019.

[32] B. Shuai, Z. Zuo, B. Wang, and G. Wang, "DAG-recurrent neural networks for scene labeling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3620–3629.

[33] Y. Chen, Z. Lin, X. Zhao, G. Wang, and Y. Gu, "Deep learning-based classification of hyperspectral data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2094–2107, Jun. 2014.

[34] X. Zhang, X. Wang, X. Tang, H. Zhou, and C. Li, "Description generation for remote sensing images using attribute attention mechanism," *Remote Sens.*, vol. 11, no. 6, p. 612, Mar. 2019.

[35] J. M. Haut, M. E. Paoletti, J. Plaza, A. Plaza, and J. Li, "Visual attention-driven hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 8065–8080, Oct. 2019.

[36] M. E. Paoletti, J. M. Haut, J. Plaza, and A. Plaza, "Deep learning classifiers for hyperspectral imaging: A review," *ISPRS J. Photogramm. Remote Sens.*, vol. 158, pp. 279–317, Dec. 2019.

[37] Y. Chen *et al.*, "Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution," 2019, *arXiv:1904.05049*. [Online]. Available: http://arxiv.org/abs/1904.05049

[38] H. Sun, X. Zheng, X. Lu, and S. Wu, "Spectral–spatial attention network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3232–3245, May 2020.

[39] P. Liu, H. Zhang, and K. B. Eom, "Active deep learning for classification of hyperspectral images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 2, pp. 712–724, Feb. 2017.

[40] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 6232–6251, Oct. 2016.

[41] Y. Zhan, D. Hu, Y. Wang, and X. Yu, "Semisupervised hyperspectral image classification based on generative adversarial networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 2, pp. 212–216, Feb. 2018.

[42] J. Yue, W. Zhao, S. Mao, and H. Liu, "Spectral–spatial classification of hyperspectral images using deep convolutional neural networks," *Remote Sens. Lett.*, vol. 6, no. 6, pp. 468–477, Jun. 2015.

[43] L. Jiao, M. Liang, H. Chen, S. Yang, H. Liu, and X. Cao, "Deep fully convolutional network-based spatial distribution prediction for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 10, pp. 5585–5599, Oct. 2017.

[44] Z. Niu, W. Liu, J. Zhao, and G. Jiang, "DeepLab-based spatial feature extraction for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 2, pp. 251–255, Feb. 2019.

[45] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[46] X. Yang, Y. Ye, X. Li, R. Y. K. Lau, X. Zhang, and X. Huang, "Hyperspectral image classification with deep learning models," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 9, pp. 5408–5423, Sep. 2018.

[47] L. Mou, P. Ghamisi, and X. Xiang Zhu, "Unsupervised spectral–spatial feature learning via deep residual Conv-Deconv network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 1, pp. 391–406, Jan. 2018.

[48] R. Hang, Q. Liu, D. Hong, and P. Ghamisi, "Cascaded recurrent neural networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5384–5394, Aug. 2019.

[49] L. Mou and X. X. Zhu, "Learning to pay attention on spectral domain: A spectral attention module-based convolutional network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 110–122, Jan. 2020.

[50] M. Zhang, W. Li, and Q. Du, "Diverse region-based CNN for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 2623–2634, Jun. 2018.

[51] L. Mou, P. Ghamisi, and X. X. Zhu, "Deep recurrent neural networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3639–3655, Jul. 2017.

[52] Z. Zhong, J. Li, Z. Luo, and M. Chapman, "Spectral–spatial residual network for hyperspectral image classification: A 3-D deep learning framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 847–858, Feb. 2018.
[53] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp. 116–131, 2018.
[54] G. Cheng, Z. Li, J. Han, X. Yao, and L. Guo, "Exploring hierarchical convolutional features for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 11, pp. 6712–6722, Nov. 2018.
[55] C. Deng, Y. Xue, X. Liu, C. Li, and D. Tao, "Active transfer learning network: A unified deep joint spectral–spatial feature learning model for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 3, pp. 1741–1754, Mar. 2019.
[56] C.-I. Chang, T.-L. Sun, and M. L. Althouse, "Unsupervised interference rejection approach to target detection and classification for hyperspectral imagery," *Opt. Eng.*, vol. 37, no. 3, p. 735, Mar. 1998.
[57] J. Zhu, L. Fang, and P. Ghamisi, "Deformable convolutional neural networks for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 8, pp. 1254–1258, Aug. 2018.
[58] W. Song, S. Li, L. Fang, and T. Lu, "Hyperspectral image classification with deep feature fusion network," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 6, pp. 3173–3184, Jun. 2018.
[59] S. Wang, X. Hou, and X. Zhao, "Automatic building extraction from high-resolution aerial imagery via fully convolutional encoder-decoder network with non-local block," *IEEE Access*, vol. 8, pp. 7313–7322, 2020.

Xu Tang (Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees in electronic circuit and system from Xidian University, Xi'an, China, in 2007, 2010, and 2017, respectively.

From 2015 to 2016, he was a joint Ph.D. student along with Prof. W. J. Emery at the University of Colorado at Boulder, Boulder, CO, USA. He is currently an Associate Professor with the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, School of Artificial Intelligence, Xidian University. He is also a Hong Kong Scholar with the Hong Kong Baptist University. His research interests include remote sensing image content-based retrieval and reranking, hyperspectral image processing, remote sensing scene classification, and object detection.

Fanbo Meng (Graduate Student Member, IEEE) received the B.Eng. degree in oil engineering from Yanshan University, Qinhuangdao, China, in 2018. He is pursuing the master's degree with the School of Artificial Intelligence, Xidian University, Xi'an, China.

His research interests include machine learning and hyperspectral image processing.

Xiangrong Zhang (Senior Member, IEEE) received the B.S. and M.S. degrees from the School of Computer Science, Xidian University, Xi'an, China, in 1999 and 2003, respectively, and the Ph.D. degree from the School of Electronic Engineering, Xidian University, in 2006.

From January 2015 to March 2016, she was a Visiting Scientist with Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA. She is a Professor with the Key Laboratory of Intelligent Perception and Image Understanding of the Ministry of Education, Xidian University. Her research interests include pattern recognition, machine learning, and remote sensing image analysis and understanding.

Yiu-Ming Cheung (Fellow, IEEE) received the Ph.D. degree from the Department of Computer Science and Engineering, Chinese University of Hong Kong, Hong Kong.

He is a Full Professor with the Department of Computer Science, Hong Kong Baptist University, Hong Kong. His research interests include machine learning, pattern recognition, visual computing, and optimization.

Dr. Cheung is a fellow of the IET, the British Computer Society, and RSA, and a Distinguished Fellow of IETI. He is the Founding Chair of the Computational Intelligence Chapter of the IEEE Hong Kong Section, and the Chair of the Technical Committee on Intelligent Informatics of the IEEE Computer Society. He serves as an Associate Editor for the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, the IEEE TRANSACTIONS ON CYBERNETICS, *Pattern Recognition*, *Knowledge and Information Systems*, and *Neurocomputing*, to name a few.

Jingjing Ma (Member, IEEE) received the B.S. and Ph.D. degrees from Xidian University, Xi'an, China, in 2004 and 2012, respectively.

She is an Associate Professor with the Key Laboratory of Intelligent Perception and Image Understanding, Ministry of Education, Xidian University. Her research interests include computational intelligence and image understanding.

Fang Liu (Member, IEEE) was born in China, in 1990. She received the B.S. degree in information and computing science from Henan University, Kaifeng, China, in 2012, and the Ph.D. degree in intelligent information processing from Xidian University, Xi'an, China, in 2018.

She is a Lecturer with the Nanjing University of Science and Technology, Nanjing, China. Her research interests include deep learning, object detection, polarimetric SAR image classification, and change detection.

Licheng Jiao (Fellow, IEEE) received the B.S. degree in high voltage from Shanghai Jiao Tong University, Shanghai, China, in 1982, and the M.S. and Ph.D. degrees in electronic engineering from Xi'an Jiaotong University, Xi'an, China, in 1984 and 1990, respectively.

From 1984 to 1986, he was an Assistant Professor with the Civil Aviation Institute of China, Tianjin, China. From 1990 to 1991, he was a Post-Doctoral Fellow with the Key Laboratory for Radar Signal Processing, Xidian University, Xi'an, where he is the Director of the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education of China. He has authored or coauthored more than 200 scientific articles. His research interests include signal and image processing, nonlinear circuits and systems theory, wavelet theory, natural computation, and intelligent information processing.

Dr. Jiao is also a member of the IEEE Xian Section Executive Committee and an Executive Committee Member of the Chinese Association of Artificial Intelligence. He is the Chairman of the Awards and Recognition Committee.